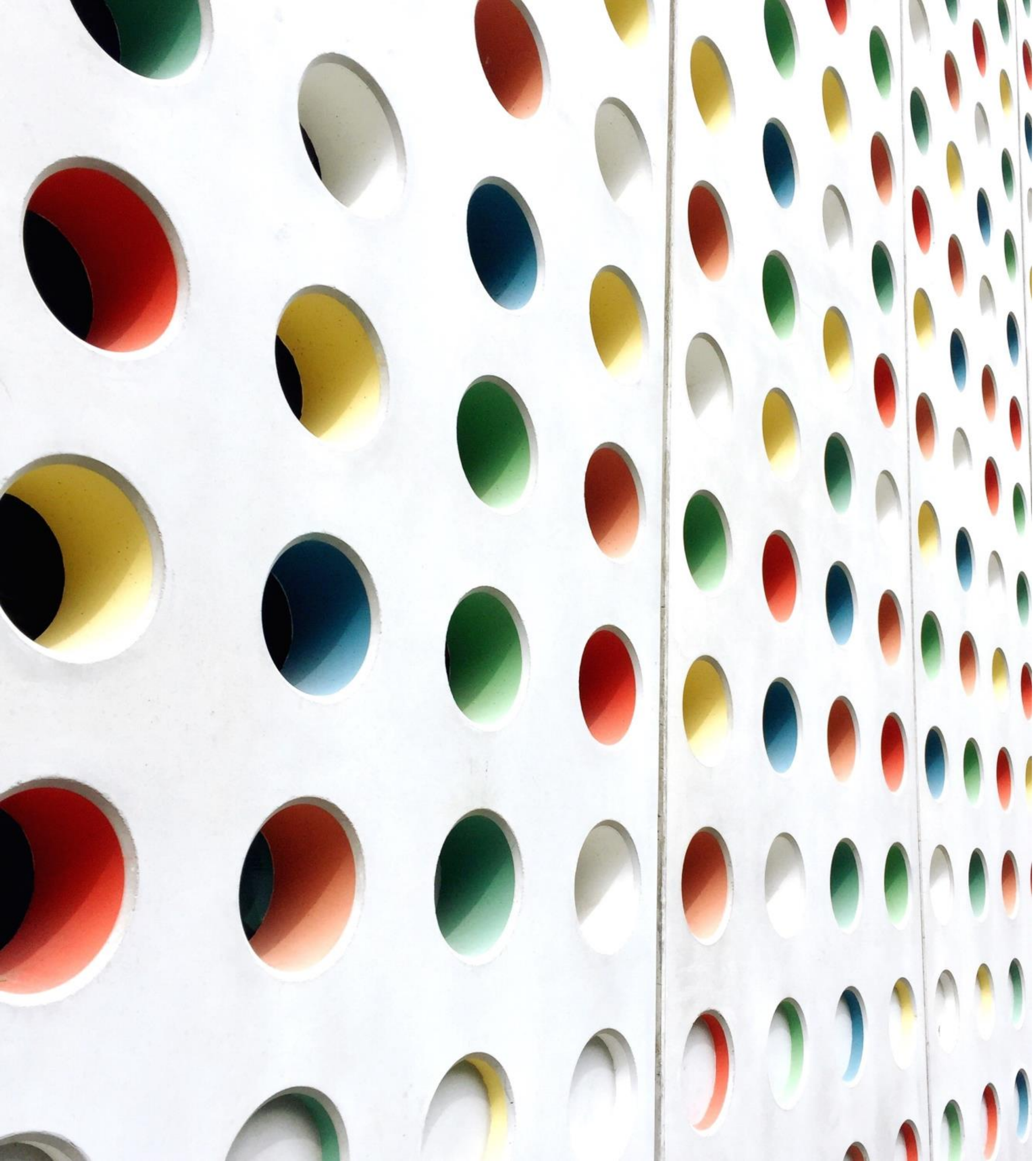


Вероjатност

Introduction – Frequentist vs Bayesian

Probability theory is nothing, but common sense reduced to calculation. – Pierre Laplace

Probability that a "fair" coin will land heads is 0.5

(At least) Two interpretations of probability:

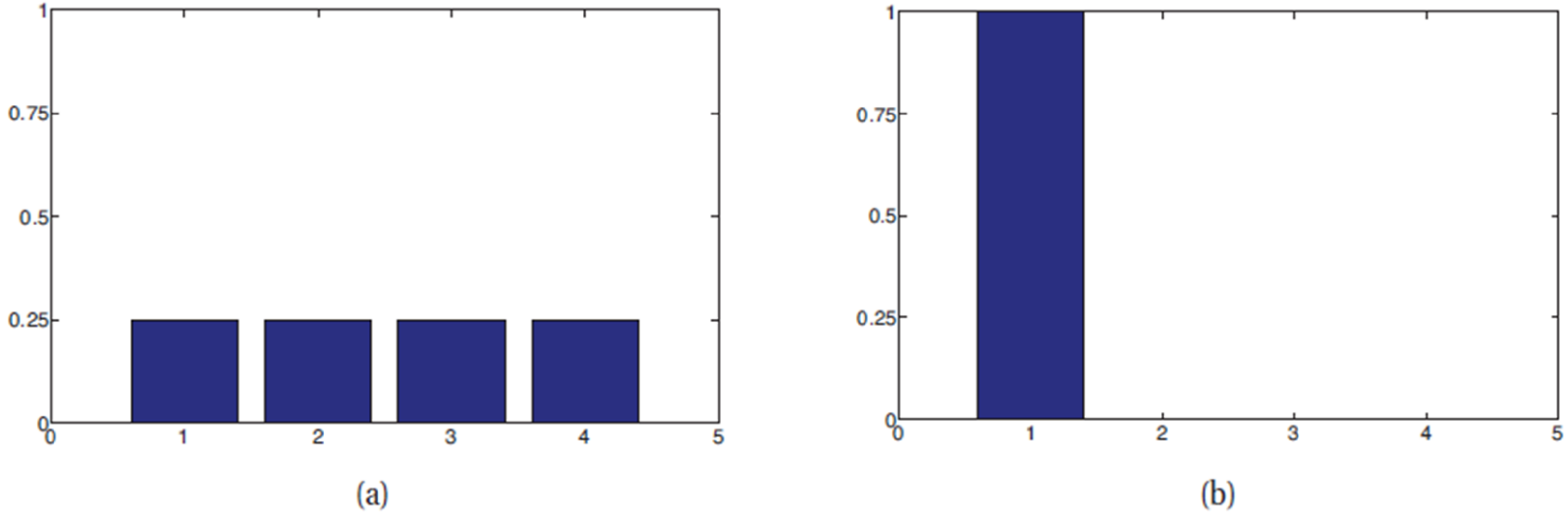
- Frequentist interpretation - probabilities represent long run frequencies of events.
  If we flip the coin many times, we expect it to land heads about half the time
- Bayesian interpretation - probability is used to quantify our **uncertainty** about something (fundamentally related to information rather than repeated trials)
  we believe the coin is equally likely to land heads or tails on the next toss

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies

Discrete Random Variables

Discrete random variable X

- Can take any value from a finite or countably infinite set $\mathcal{X}$
- Probability that X = x is denoted by $p(X=x)$ or $p(x)$, where $p()$ is a probability mass function – pmf
- (a) uniform distribution, and (b) degenerate distribution



(a)

(b)

Fundamental Rules

Probability of a union of two events

$$\begin{aligned} p(A \vee B) &= p(A) + p(B) - p(A \wedge B) \\ &= p(A) + p(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \end{aligned}$$

Joint probabilities

$$p(A, B) = p(A \wedge B) = p(A|B)p(B)$$

Marginal distribution (sum rule)

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

Conditional probability

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

Bayes Rule

Bayes rule or **Bayes theorem**

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Example: Generative classifiers

- Specifies how to generate the data using the **class conditional density** $p(x|y = c)$ and the **class prior** $p(y = c)$.

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(y = c|\boldsymbol{\theta})p(\mathbf{x}|y = c, \boldsymbol{\theta})}{\sum_{c'} p(y = c'|\boldsymbol{\theta})p(\mathbf{x}|y = c', \boldsymbol{\theta})}$$

Example: medical diagnosis

As an example of how to use this rule, consider the following medical diagnosis problem. Suppose you are a woman in your 40s, and you decide to have a medical test for breast cancer called a **mammogram**. If the test is positive, what is the probability you have cancer? That obviously depends on how reliable the test is. Suppose you are told the test has a **sensitivity** of 80%, which means, if you have cancer, the test will be positive with probability 0.8. The prior probability of having breast cancer, which fortunately is quite low is:

$$p(y = 1) = 0.004$$

We also need to take into account the fact that the test may be a **false positive** or **false alarm**. Unfortunately, such false positives are quite likely (with current screening technology):

$$p(x = 1|y = 0) = 0.1$$

Example: medical diagnosis

x – has positive test

y – has cancer

$p(y{=}1) = 0.004$

$p(x{=}1, y{=}1) = 0.8$

$p(x{=}1, y{=}0) = 0.1$

$$\boxed{\begin{aligned} p(y=1|x=1) &= \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1)+p(x=1|y=0)p(y=0)} \\ &= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031 \end{aligned}}$$

Therefore, if you test positive, you only have about a 3% chance of actually having breast cancer

Прашање

Еден човек кој работи во една компанија поседува два автомобили: голем и мал. Три четвртини од времето го користи малиот автомобил за да оди на работа. Ако го земе малиот автомобил, најчесто полесно наоѓа место за паркирање, па стигнува на работа на време со веројатност од 0.9. Ако го земе големиот автомобил, тој стигнува навремено на работа со веројатност од 0.6. Ако знаеме дека едно утро стигнал на работа на време, која е веројатноста (заокружена на две децимали) да го возел малиот автомобил?

a) 0.67

b) 0.82

c) 0.68

d) 0.45

Independence and conditional independence

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

$$X \perp Y \iff p(X,Y) = p(X)p(Y)$$

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

$$X \perp Y|Z \iff p(X,Y|Z) = p(X|Z)p(Y|Z)$$

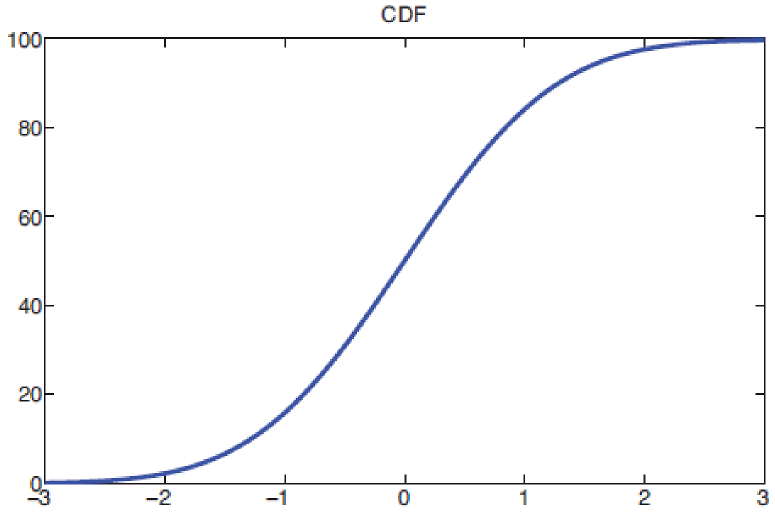
Example: the probability it will rain tomorrow (event *X*) is independent of whether the ground is wet today (event *Y*), given knowledge of whether it is raining today (event *Z*). Intuitively, this is because *Z* “causes” both *X* and *Y* , so if we know *Z*, we do not need to know about *Y* in order to predict *X* or vice versa

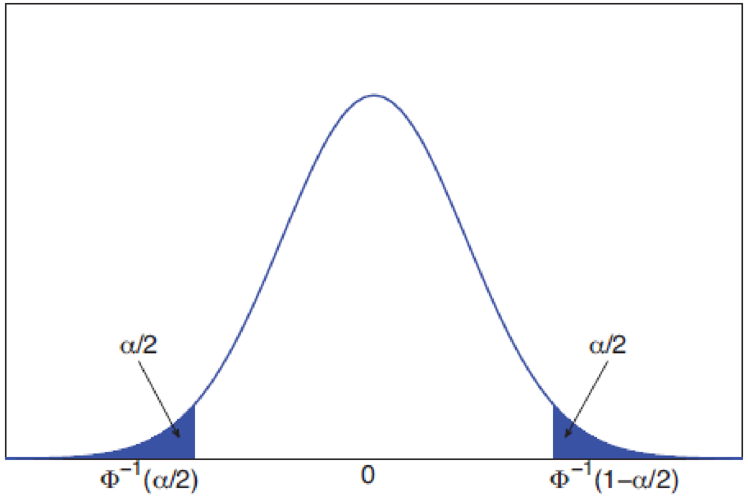
Continuous Random Variables

Continuous random variable X

- Cumulative distribution function, cdf $F(q) \triangleq p(X \leq q)$
- Probability density function, pdf $f(x) = \frac{d}{dx}F(x)$
- Given a pdf we can compute the probability of a continuous variable being in a finite interval as follows:

$$P(a < X \leq b) = \int_a^b f(x)dx$$





Measures

Quantiles

- If F is the cdf of X, then $\mathrm{F}^{-1}(\alpha)$ is the value of x_α such that $P(X \leq x_\alpha) = \alpha$;
- This is called the α quantile of F
- $\mathrm{F}^{-1}(0.5)$ – **median** (half of the probability mass on the left, and half on the right)
- $\mathrm{F}^{-1}(0.25)$ and $\mathrm{F}^{-1}(0.75)$ – lower and upper **quartiles**

Mean (expected value) $\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x\ p(x)$ $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\ p(x)dx$

Variance (measure of the "spread" of a distribution)

- Standard Deviation

$$\mathrm{std}[X] \triangleq \sqrt{\mathrm{var}[X]}$$

$$\begin{aligned}\mathrm{var}[X] &\triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx \\ &= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu \int xp(x)dx = \mathbb{E}\left[X^2\right] - \mu^2\end{aligned}$$

Some common discrete distributions

The **binomial** and **Bernoulli** distributions, $X \sim \mathrm{Bin}(n, \theta)$ (Bernoulli is a special case of a Binomial distribution with n=1)

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k} \qquad \mathrm{mean} = \theta, \quad \mathrm{var} = n\theta(1-\theta)$$

The **multinomial** and **multinoulli** (**categorical**) distributions

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

The **Poisson** distribution, $X \sim \mathrm{Poi}(\lambda)$ – model for count of events in fixed time

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

The **empirical** distribution

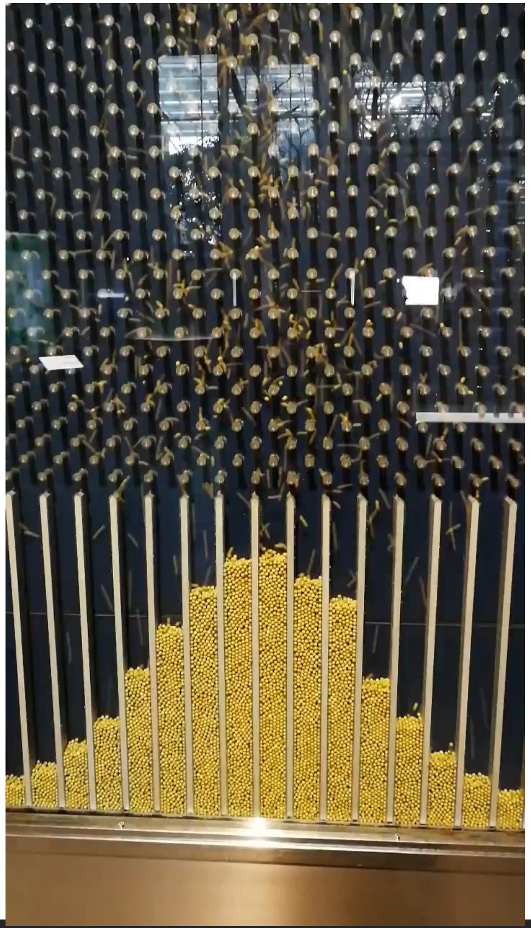
$$p_{\mathrm{emp}}(A) \triangleq \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}(A)$$

Some common continuous distributions – Gaussian distribution

The **Gaussian** (normal) distribution, $X \sim N(\mu, \sigma^2)$. Its pdf is

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

- Here $\mu = \mathrm{E}\,[\mathrm{X}]$ is the mean, and $\sigma^2 = \mathrm{var}\,[\mathrm{X}]$ is the variance
- If $X \sim N(0, 1)$ we say X follows a standard normal distribution
- Precision of Gaussian $\lambda = 1/\sigma^2$ (high precision means a narrow distribution)
- The Gaussian distribution is the most widely used distribution in statistics. There are several reasons for this:
  - It has two parameters which are easy to interpret, and which capture some of the most basic properties of a distribution, namely its mean and variance.
  - The central limit theorem tells us that sums of independent random variables have an approximately Gaussian distribution, making it a good choice for modeling residual errors or "noise"
  - The Gaussian distribution makes the least number of assumptions, subject to the constraint of having a specified mean and variance
  - It has a simple mathematical form, which results in easy to implement, but often highly effective methods



Some common continuous distributions

One problem with the Gaussian distribution is that it is sensitive to outliers. A more robust distribution is the **Student t distribution**

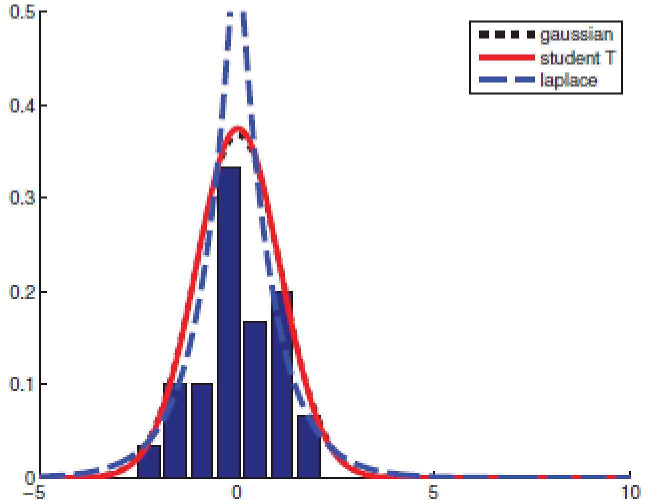
$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)} \qquad \text{mean} = \mu, \text{mode} = \mu,\ \text{var} = \frac{\nu\sigma^2}{(\nu-2)}$$

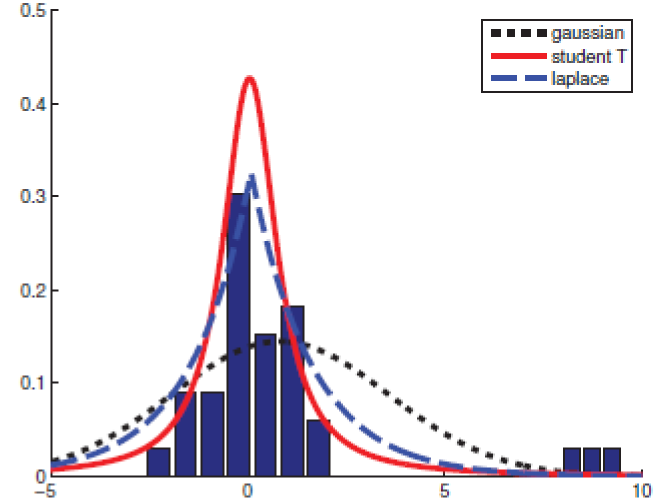
- where μ is the mean, $\sigma^2 > 0$ is the scale parameter, and $\nu > 0$ is called the **degrees of freedom**
- The mean is defined if $\nu > 1$ and the variance if $\nu > 2$

The **Laplace distribution** (double sided exponential distribution)

$$\text{Lap}(x|\mu,b) \triangleq \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \qquad \text{mean} = \mu,\ \text{mode} = \mu,\ \text{var} = 2b^2$$

- where μ is a location parameter and $b > 0$ is a scale parameter





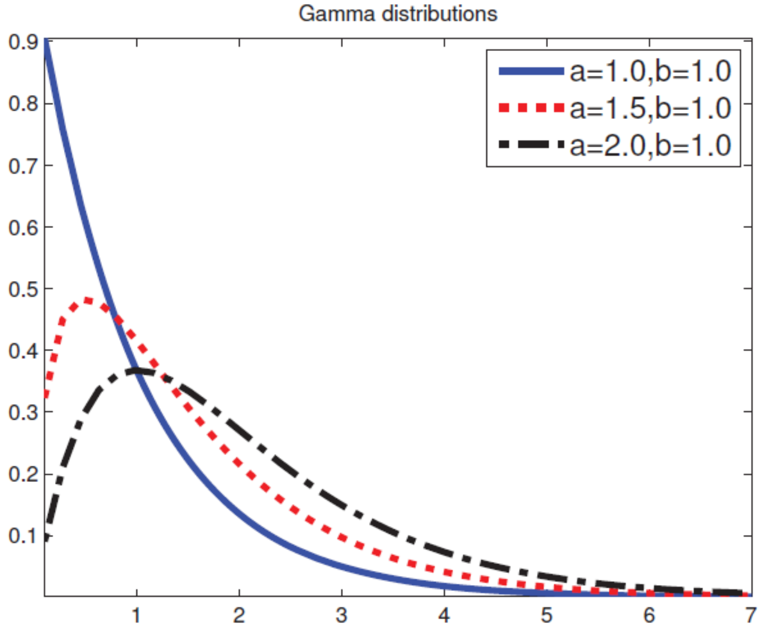
Some common continuous distributions

The **gamma** distribution, $X \sim \mathrm{Ga}(a, b)$ – flexible distribution for positive real valued rvs

$$\mathrm{Ga}(T|\mathrm{shape}=a, \mathrm{rate}=b) \triangleq \frac{b^a}{\Gamma(a)} T^{a-1} e^{-Tb} \quad \mathrm{mean} = \frac{a}{b}, \ \mathrm{mode} = \frac{a-1}{b}, \ \mathrm{var} = \frac{a}{b^2}$$

$$\Gamma(x) \triangleq \int_0^\infty u^{x-1} e^{-u} du$$

- **Exponential** distribution, $\mathrm{Expon}(x|\lambda) = \mathrm{Ga}(x|1, \lambda)$
- **Erlang** distribution, $\mathrm{Erlang}(x|\lambda) = \mathrm{Ga}(x|a, \lambda)$, (a is integer, common to fix a=2) – models amount of time for certain number of event occurrences
- **Chi-squared** distribution, $\chi^2(x|v) = \mathrm{Ga}(x|v/2, 1/2)$



Some common continuous distributions

The **beta** distribution has support over the interval [0, 1] and is defined as follows

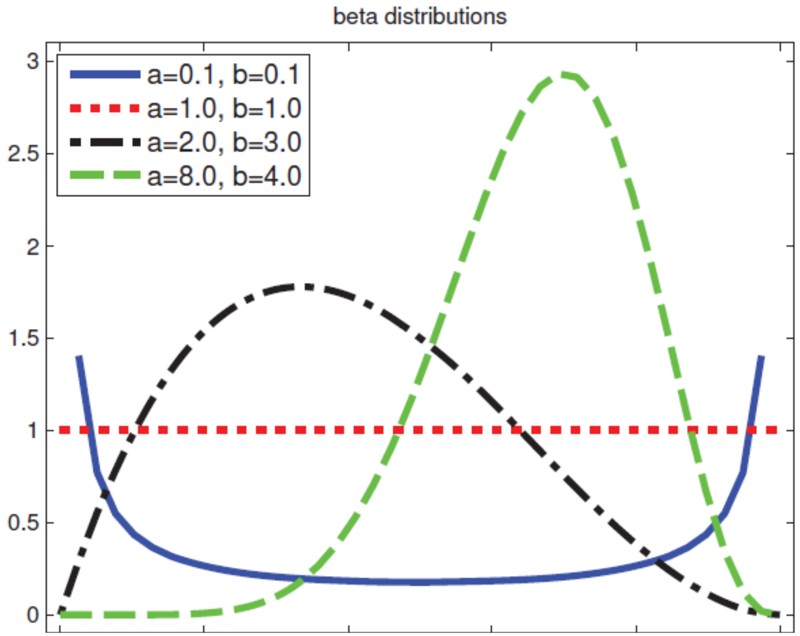
$$\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \qquad B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

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

If a = b = 1, we get the uniform distribution.

If a and b are both less than 1, we get a bimodal distribution with “spikes” at 0 and 1;

If a and b are both greater than 1, the distribution is unimodal.



Some common continuous distributions

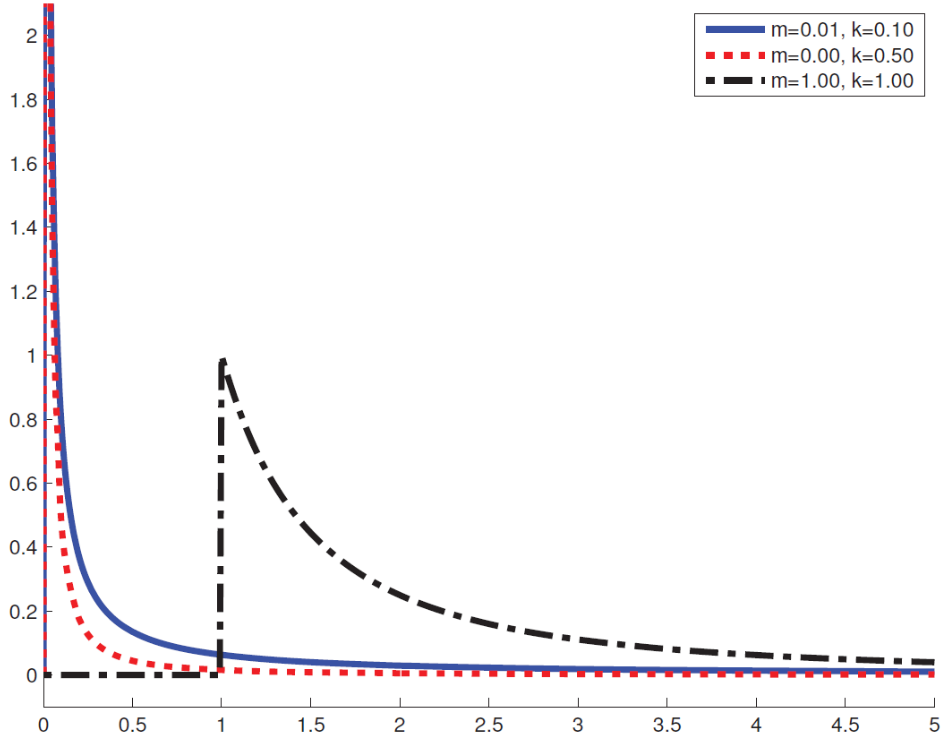
The **Pareto** distribution

- used to model distribution of quantities that exhibit **long tails (heavy tails)**

$$\text{Pareto}(x|k,m) = km^k x^{-(k+1)}\mathbb{I}(x \geq m)$$

$$\text{mean} = \frac{km}{k-1} \text{ if } k > 1\ , \quad \text{mode} = m, \quad \text{var} = \frac{m^2 k}{(k-1)^2(k-2)} \text{ if } k > 2$$

- Examples include:
  - **inequality in wealth distribution** (80% of the country's wealth was concentrated in the hands of only 20% of the population)
  - Company revenues



Прашање

Норамалната дистрибуција е симетрична во однос на:

a) Варијансата

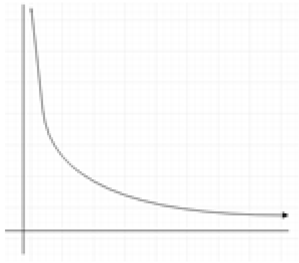
b) Средната вредност

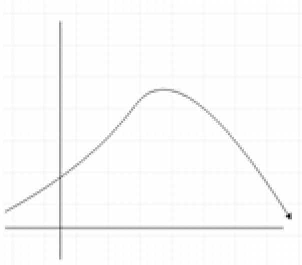
c) Стандардната варијација

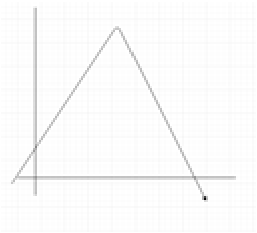
d) Коваријансата

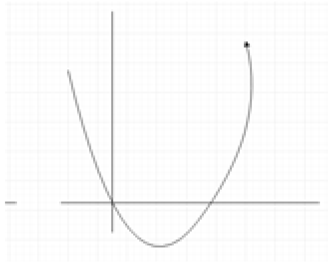
Прашање

Која од следните распределби претставува Gamma distribution?









a) Број 1

b) Број 2

c) Број 3

d) Број 4

Joint probability distributions

Building joint distributions on multiple related random variables

A joint probability distribution has the form $p(x_1, \ldots, x_D)$ for a set of $D > 1$ variables, and models the relationships between the variables.

If all the variables are discrete, we can represent the joint distribution as a big multi-dimensional array, with one variable per dimension.

However, the number of parameters needed to define such a model is $O(K^D)$, where K is the number of states for each variable.

We can define high dimensional joint distributions using fewer parameters by making conditional independence assumptions.

Covariance and correlation

The **covariance** between two rv's X and Y measures the degree to which they are linearly related

$$\mathrm{cov}\left[X,Y\right] \triangleq \mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \qquad -\infty \le \mathrm{cov}[X,Y] \le \infty.$$

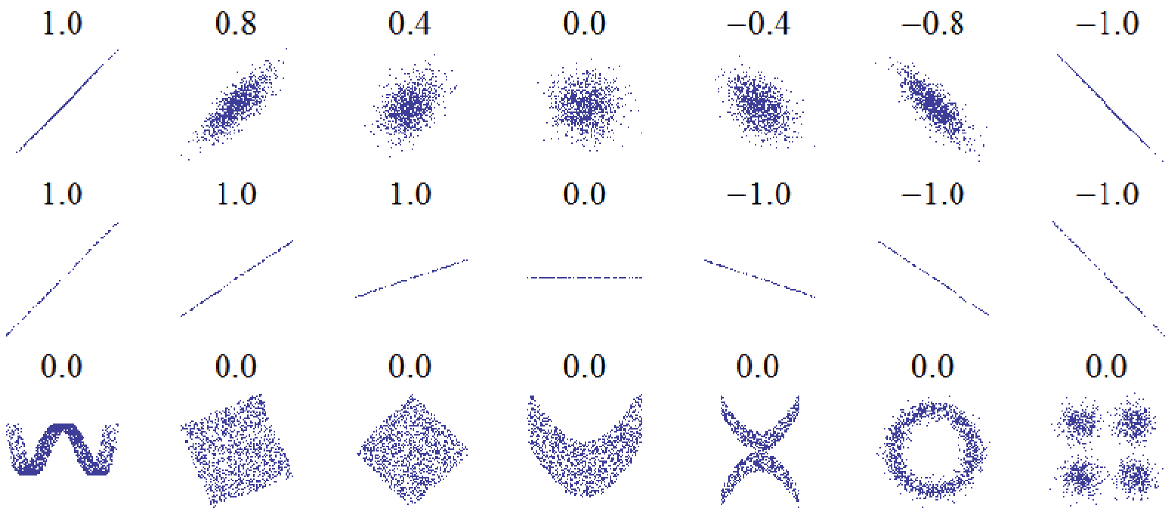
The (**Pearson**) **correlation coefficient** between X and Y is defined as

$$\mathrm{corr}\left[X,Y\right] \triangleq \frac{\mathrm{cov}\left[X,Y\right]}{\sqrt{\mathrm{var}\left[X\right]\mathrm{var}\left[Y\right]}}$$

where $-1 \le \mathrm{corr}[X,Y] \le 1$.

Independence implies no correlation

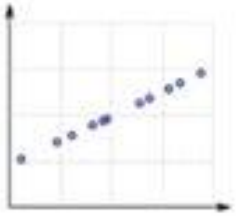
- Uncorrelated does not imply independent

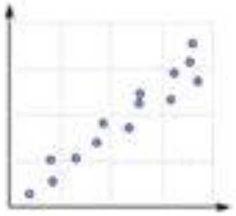


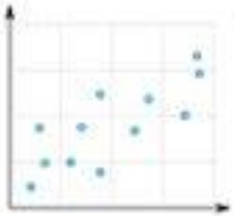
Прашање

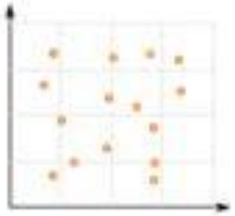
Да претпоставиме дека ви се дадени 7 Scatter plots 1-7 (лево надесно) и сакате да ги споредите Пирсоновите корелациски коефициенти помеѓу променливите на секој график. Што од наведеното е во правилен редослед?

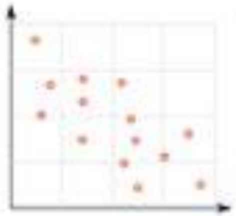
1. 1<2<3<4
2. 1>2>3>4
3. 7<6<5<4
4. 7>6>5>4

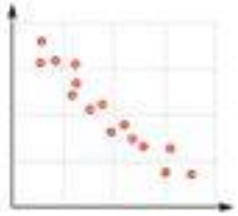

1.

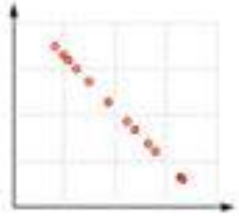

2.


3.


4.


5.


6.

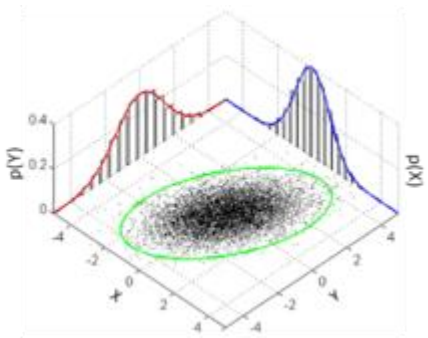

7.

a) 1 и 3

b) 2 и 3

c) 1 и 4

d) 2 и 4

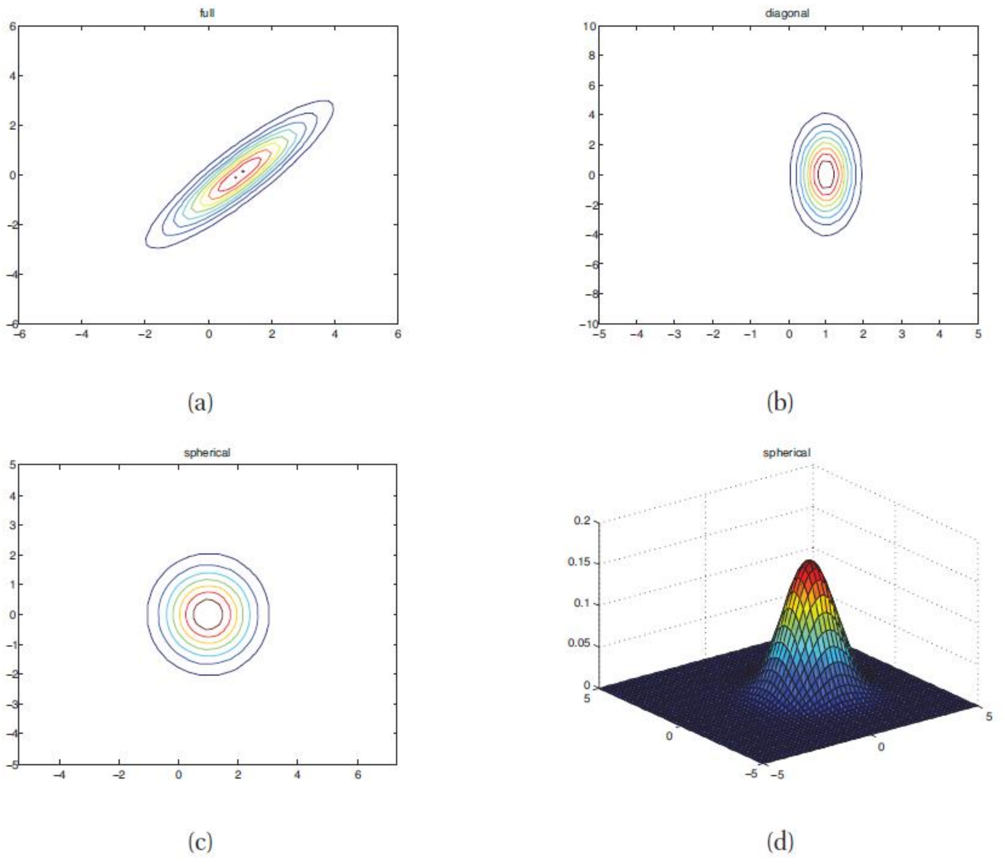


Multivariate distributions - Gaussian

The multivariate Gaussian (MVN)

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

- where $\boldsymbol{\mu}$ where is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix, $\boldsymbol{\Lambda}$ is the precision matrix, $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$
- Different kinds of covariance matrices
  - (a) Full covariance matrix – $D(D+1)/2$ elements
  - (b) Diagonal covariance matrix – D elements
  - (c, d) Isotropic covariance matrix, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_D$ - one free parameter



(a)

(b)

(c)

(d)

Multivariate distributions

Multivariate Student t distribution

$$\mathcal{T}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) \quad = \quad \frac{\Gamma(\nu/2 + D/2)}{\Gamma(\nu/2)} \frac{|\boldsymbol{\Sigma}|^{-1/2}}{\nu^{D/2}\pi^{D/2}} \times \left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-(\frac{\nu+D}{2})}$$

- Σ is called the scale matrix (it is not exactly the covariance matrix)
- Has fatter tails than a Gaussian (the smaller the ν, the fatter the tails. As $\nu \longrightarrow \infty$, the distribution tends towards a Gaussian.)

Transformation of Random Variables

If $x \sim p()$ is some random variable, and $y=f(x)$, what is the distribution of y?

Linear transformations

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\mu + \mathbf{b}$$

$$\mathbb{E}\left[\mathbf{a}^T\mathbf{x} + b\right] = \mathbf{a}^T\boldsymbol{\mu} + b$$

Where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$. This is called the linearity of expectation

$$\text{cov}[\mathbf{y}] = \text{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

$$\text{var}[y] = \text{var}\left[\mathbf{a}^T\mathbf{x} + b\right] = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}$$

$$\boldsymbol{\Sigma} = \text{cov}[\mathbf{x}]$$

Note, however, that the mean and covariance only completely define the distribution of y if x is Gaussian. In general, we must use the following techniques to derive the full distribution of y, as opposed to just its first two moments.

Transformation of Random Variables

General transformations

- If X is a discrete rv, we can derive the pmf for y by simply summing up the probability mass for all the x's such that $f(x) = y$:

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

- For example, if $f(X) = 1$ if X is even and $f(X) = 0$ otherwise, and $p_x(X)$ is uniform on the set {1, . . ., 10}, then $p_y(1) = \sum_{x\in\{2,4,6,8,10\}} p_x(x) = 0.5$, and $p_y(0) = 0.5$ similarly
- If X is continuous, we cannot use $p(x)$ since it is a density, not a pmf, and we cannot sum up densities. Therefore, we use the cdf.

$$P_y(y) \triangleq P(Y \leq y) = P(f(X) \leq y) = P(X \in \{x|f(x) \leq y\})$$

$$P_y(y) = P(f(X) \leq y) = P(X \leq f^{-1}(y)) = P_x(f^{-1}(y)) \qquad x = f^{-1}(y)$$

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

- For example, suppose $X \sim U(-1,1)$, and $Y = X^2$. Then, $p_y(y) = \frac{1}{2}\frac{1}{2}y^{-\frac{1}{2}}$

Central Limit Theorem

Now consider N random variables with pdf's (not necessarily Gaussian) $p(x_i)$, each with mean μ and variance σ^2

We assume each variable is **independent and identically distributed** or **iid** for short

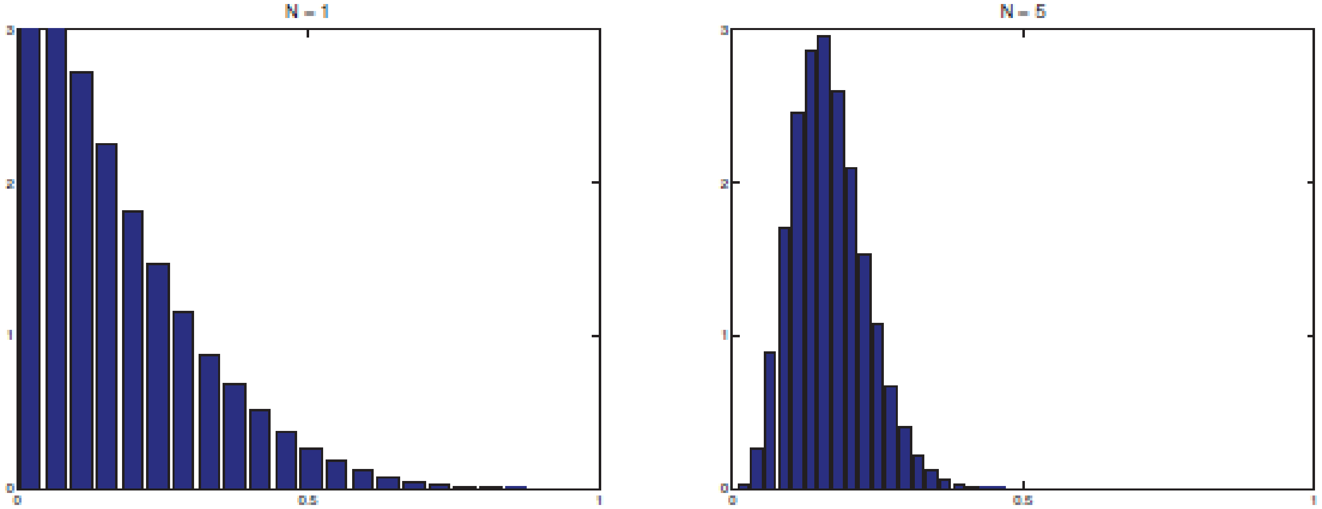
Let $S_N = \Sigma X_i$ be the sum of the rv's. This is a simple but widely used transformation of rv's.

As N increases, the distribution of this sum converges to the normal. This is called the **central limit theorem**.

An example for $x_i \sim \mathrm{Beta}(1, 5)$:

a) N=1

b) N=5



(a) (b)

Задача

Генерирајте 1000 податоци кои припаѓаат на стандардна нормална дистрибуција (користејќи функции во Python). Понатаму, претпоставете дека има вкупно 5 вакви променливи. Применете ја central limit теоремата. Што се добива како резултат за сумата на случајните променливи?

a) Сумата има Гаусова дистрибуција со средна вредност 0 и стандардна девијација 1

b) Сумата има Гаусова дистрибуција со средна вредност 0 и стандардна девијација $\sqrt{5}$

c) Сумата има Бета дистрибуција со параметри 1 и 5

d) Не може да се одреди дистрибуцијата на сумата

Monte Carlo approximation

A way of computing the distribution of a function of a rv

First, we generate S samples from the distribution, call them $x_1, \ldots, x_S$.

Given the samples, we can approximate the distribution of $f(X)$ by using the empirical distribution of $\{f(x_s)\}$.

We can use Monte Carlo to approximate the expected value of any function of a random variable. We simply draw samples, and then compute the arithmetic mean of the function applied to the samples.

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)$$

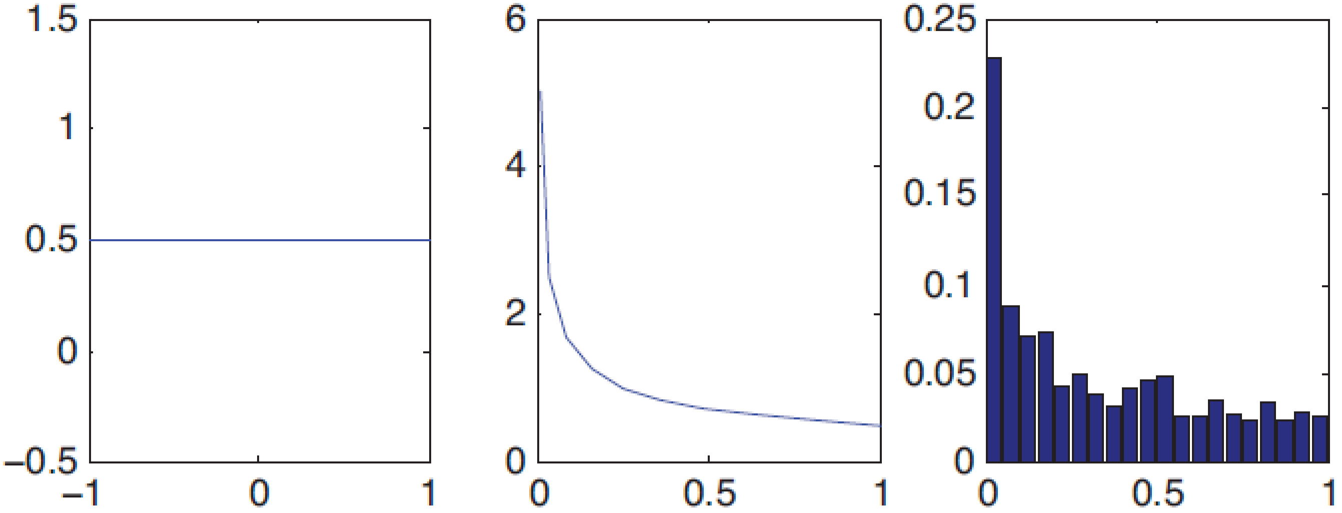
By varying the function $f()$, we can approximate many quantities of interest, such as

$$\overline{x} = \frac{1}{S}\sum_{s=1}^{S} x_s \to \mathbb{E}[X] \qquad \frac{1}{S}\sum_{s=1}^{S}(x_s - \overline{x})^2 \to \text{var}[X] \qquad \frac{1}{S}\#\{x_s \leq c\} \to P(X \leq c)$$

Example: change of variables, the MC way

An alternative (and simpler) approach to analytically compute the distribution of a function of a random variable, $y = f(x)$ is to use a Monte Carlo approximation.

For example, suppose $x \sim \mathrm{Unif}(-1, 1)$ and $y = x^2$. We can approximate $p(y)$ by drawing many samples from $p(x)$, squaring them, and computing the resulting empirical distribution.



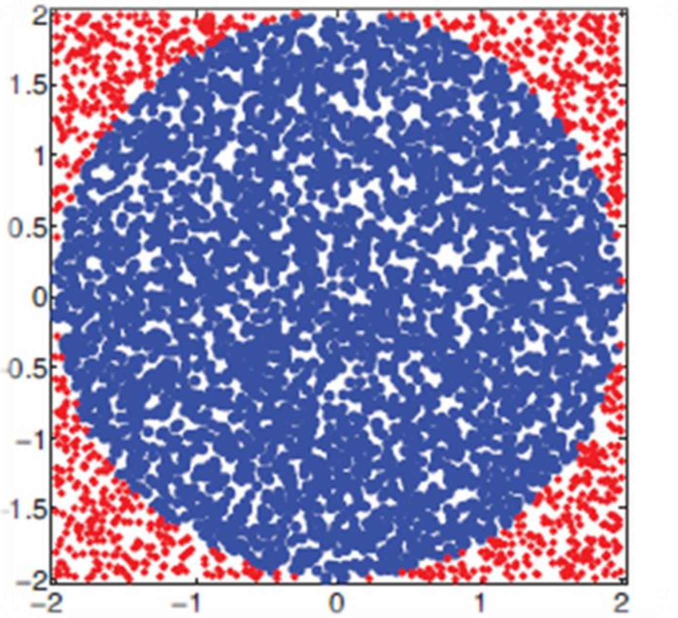
Example: estimating π by Monte Carlo integration

Suppose we want to estimate π. We know that the area of a circle with radius r is πr^2, but it is also equal to the following definite integral:

$$I = \int_{-r}^{r} \int_{-r}^{r} \mathbb{I}(x^2 + y^2 \leq r^2) dx dy$$

Hence $\pi = I/(r^2)$

Let $f(x, y) = I(x^2 + y^2 \leq r^2)$ be an indicator function that is 1 for points inside the circle, and 0 outside, and let $p(x)$ and $p(y)$ be uniform distributions on $[-r, r]$



Example:

- *Area of square = 16*
- *Area of circle =* $\pi * r^2 = \pi * 4$
- $\frac{area\ of\ circle}{area\ of\ square} = \frac{\#\ points\ inside\ circle}{\#\ points\ inside\ square} = \frac{4*\pi}{16} => \pi = 4 * \frac{\#\ points\ inside\ circle}{\#\ points\ inside\ square}$
- We find π = 3.1416 with standard error 0.09

Information theory

Information theory is concerned with representing data in a compact fashion (a task known as **data compression** or **source coding**), as well as with transmitting and storing it in a way that is robust to errors (a task known as **error correction** or **channel coding**).

Compactly representing data requires allocating short codewords to highly probable bit strings, and reserving longer codewords to less probable bit strings (as in natural language).

In both cases, we need a model that can predict which kinds of data are likely and which unlikely, which is also a central problem in machine learning.

Entropy

The **entropy** of a random variable X with distribution $p()$, denoted by $H(X)$ or sometimes $H(p)$, is a measure of its uncertainty.

In particular, for a discrete variable with K states, it is defined by

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

Uniform distribution – maximum entropy

Delta-function – minimum entropy

How much information do you gain when you witness an event of a rv?

- A degree of surprise

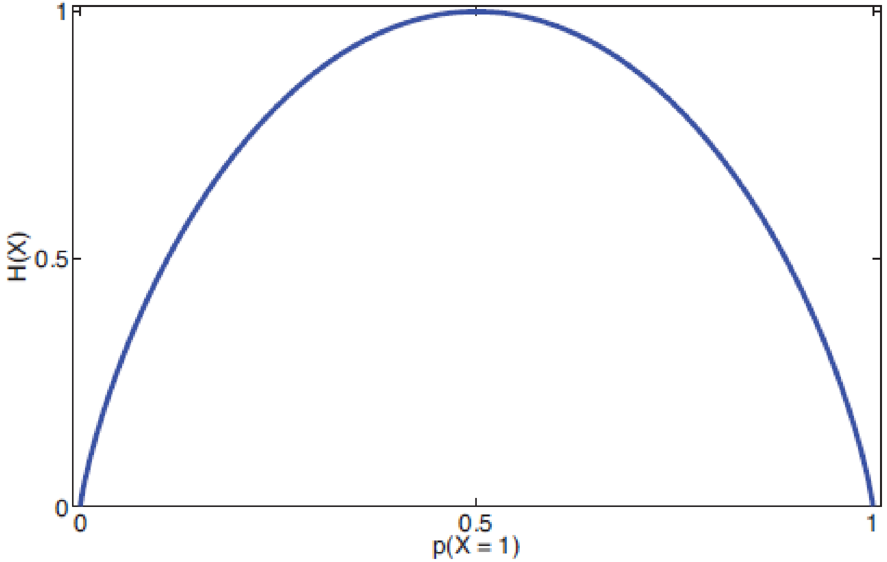
Entropy

For binary random variables $X \in \{0,1\}$, we can write $p(X=1)=\theta$ and $p(X=0)=1-\theta$

We have a **binary entropy function**,

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

Entropy of a Bernoulli rv as a function of θ



Прашање

Дали доколку една променлива има униформна дистрибуција, ентропијата ќе биде поголема отколку за друга дистрибуција каде p=[0.25, 0.25, 0.2, 0.15, 0.15]?

a) Да

b) Не, ќе биде помала

c) Не, ќе биде иста

d) Не може да се каже

Прашање

Колку ќе биде ентропијата на дискретна променлива $X \epsilon \{1, \dots, K\}$ со униформна дистрибуција?

a) $\log_2 \text{К}$

b) $-\log_2 \text{К}$

c) $\frac{1}{\text{К}} \log_2 \text{К}$

d) $-\frac{1}{\text{К}} \log_2 \text{К}$

KL divergence

One way to measure the dissimilarity of two probability distributions, p and q, is known as the **Kullback-Leibler divergence** (**KL divergence**) or **relative entropy,** which for discrete rvs is defined as

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

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

$\mathbb{H}\left(p, q\right)$ is called cross entropy

The KL divergence is the average number of *extra* bits needed to encode the data, due to the fact that we used distribution q to encode the data instead of the true distribution p.

The "extra number of bits" interpretation should make it clear that $\mathrm{KL}(p||q) \geq 0$, and that the KL is only equal to zero *iff* $q = p$.

Mutual information

Consider two random variables, X and Y. Suppose we want to know how much knowing one variable tells us about the other.

We can use correlation coefficient, but it is only for real-valued random variables, and it is a limited measure of distance

One general approach is to determine how similar the joint distribution $p(X, Y)$ is to the factored distribution $p(X)p(Y)$.

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

MI is zero iff the variables are independent

Thus, we can interpret the MI between X and Y as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X.

Materials

Kevin P. Murphy - Machine Learning A Probabilistic Perspective, Chapter 2