Введение в компьютерный и интеллектуальный анализ данных

0. Введение.
Общие сведения.

- Объем курса – 18 часов лекции
  16 часов лабораторные занятия
- Лабораторные занятия проводятся в классе ПЭВМ и выполняются в среде пакета R
- Форма отчетности – зачет
- Лектор – Воротницкая Татьяна Ивановна

0. Введение.
Что такое компьютерный анализ данных

- **Компьютерный анализ данных** - научное направление, объединяющее вероятностно-статистические, логико-алгебраические, графические, другие модели, а также алгоритмы, программные средства обработки и анализа эмпирических данных с целью получения научно-обоснованных выводов и принятия решений относительно исследуемых объектов

0. Введение.
Основные разделы

- **Статистический анализ данных** (Statistical Data Analysis – SDA)
- **Интеллектуальный анализ данных** (Data Mining или Knowledge Discovery in Database - KDD)
- **Анализ больших данных** (Big Data Analysis - BDA)

0. Введение.
Литература.

- Ширяев А.Н. Вероятность. Москва, 1980.
- Вентцель Е.С. Теория вероятностей: Учеб. для вузов. — 6-е изд. стер. — М.: Высш. шк., 1999.
- Колмогоров А.Н. Основные понятия теории вероятностей. Москва, 1936.
- Хацкевич Г.А. Статистика. Описательный подход / Г.А. Хацкевич. – Минск: НИУП. – 2002.
- А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP
- Елисеева И.И. Общая теория статистики / И.И. Елисеева, М.М. Юзбашев. – М. – 1996.
- Тюрин Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров
- Torgo L. Data Mining with R: learning by case studies / L. Torgo - LIACC-FEP, University of Porto. – 2003.

1. Основные понятия теории вероятностей

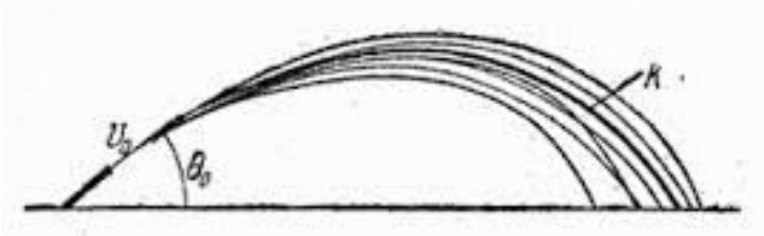
- **Теория вероятностей** - математическая наука, изучающая закономерности в случайных явлениях

- **Случайное явление** – это такое явление, которое при неоднократном воспроизведении одного и того же опыта протекает каждый раз несколько по-иному

1. Основные понятия теории вероятностей
Примеры случайных явлений

Стрельба из орудия, установленного под заданным углом к горизонту

- **Детерминированы:** начальная скорость снаряда, угол бросания, форма снаряда
- **Фактическая траектория снаряда отклоняется** за счет совокупного влияния второстепенных случайных факторов: ошибки изготовления снаряда, отклонение веса порохового заряда от номинала, неоднородность структуры и неравномерность горения заряда, ошибки установки ствола, вариации атмосферного давления и др.



1. Основные понятия теории вероятностей
Примеры случайных явлений

Взвешивание одного и того же тела на аналитических весах

- **Детерминированы:** массы тела и разновесов, геометрические форма и размеры плеч весов, значение ускорения свободного падения

- **Результаты повторных взвешиваний несколько отличаются** за счет совокупного влияния второстепенных случайных факторов: положения тела на чашке весов, случайные вибрации, ошибки отсчета показаний прибора и др.

1. Основные понятия теории вероятностей
Примеры случайных явлений

Бросание игральной кости

- **Детерминированы:** форма (куб) и распределение плотности материала (в идеале – равномерное)
- **Результаты повторных выбрасываний отличаются** за счет случайных направлений и скоростей поступательного и вращательного движений при бросании кости

1. Основные понятия теории вероятностей
Какие закономерности изучает теория вероятностей

- Теория вероятностей изучает закономерности, проявляющиеся при рассмотрении большого числа однородных случайных явлений.

- Закономерности, проявляющиеся в массе случайных явлений нивелируют, «погашают» индивидуальные особенности каждого из случайных явлений.

- **Методы теории вероятностей по природе приспособлены только для исследования массовых случайных явлений**; они не дают возможности предсказать исход отдельного случайного явления, но **дают возможность предсказать средний суммарный результат массы однородных случайных явлений**, предсказать средний исход массы аналогичных опытов, конкретный исход каждого из которых остается неопределенным, случайным.

1. Основные понятия теории вероятностей
Событие

- Под «событием» в теории вероятностей понимается всякий факт, который в результате опыта может произойти или не произойти.
- Примеры событий:
  - Появление герба при однократном бросании монеты
  - появление трех гербов при трехкратном бросании монеты;
  - попадание в цель при выстреле;
  - появление туза при вынимании карты из колоды;
  - обнаружение объекта при одном цикле обзора радиолокационной станции;
  - обрыв нити в течение часа работы ткацкого станка.
- Каждое событие обладает различной степенью возможности.
- С каждым событием можно попытаться связать некоторое число, характеризующее объективную возможность события – вероятность.
- Единица измерения вероятностей вероятность достоверного события = 1. Вероятность невозможного события = 0.

1. Основные понятия теории вероятностей
Статистическая устойчивость

- Если A – некоторое случайное событие, то доля m/n экспериментов, в которых данное событие произошло, имеет тенденцию стабилизироваться с ростом общего числа экспериментов n, приближаясь к некоторому числу p(A). Это число служит объективной характеристикой «степени возможности» произойти событию A

Пример: эксперимент по бросанию монеты.

- Случайное событие – выпадение герба
- Проведем по 10 экспериментов, в каждом из которых будем проводить n испытаний, $n=10^2$, 10^4, 10^6.
- Число выпадений герба в каждой серии обозначим m.
- В таблице показаны значения m в каждом из экспериментов и значения относительной частоты p(A)=m/n выпадений герба при различном числе испытаний

1. Основные понятия теории вероятностей
Статистическая устойчивость

Номер эксперимента	$n=10^2$		$n=10^4$		$n=10^6$	
	m	p	m	p	m	p
1	41	0,41	4985	0,4985	499558	0,499558
2	48	0,48	5004	0,5004	499952	0,499952
3	44	0,44	5085	0,5085	500114	0,500114
4	52	0,52	4946	0,4946	500064	0,500064
5	58	0,58	4978	0,4978	500183	0,500183
6	52	0,52	4985	0,4985	499533	0,499533
7	45	0,45	5012	0,5012	500065	0,500065
8	50	0,5	4931	0,4931	500317	0,500317
9	52	0,52	5016	0,5016	500449	0,500449
10	45	0,45	4973	0,4973	500704	0,500704

Очевидна стабилизация относительной частоты p(A)=m/n выпадений герба с ростом числа испытаний n, а также стремление p(A) к величине ½.

1. Основные понятия теории вероятностей. **Пространство элементарных исходов.**

- Пространством элементарных событий Ω называется множество, содержащее все возможные случайные результаты данного эксперимента, из которых в эксперименте происходит ровно один. Элементы этого множества называют элементарными исходами ω.
- События будем называть подмножества множества Ω. Говорят, что в результате эксперимента произошло событие $A \subseteq \Omega$, если в эксперименте произошел один из элементарных исходов, входящих в множество A.

1. Основные понятия теории вероятностей. **Пространство элементарных исходов.**

Пример: однократное подбрасывание игральной кости.

- Пространством элементарных событий Ω = {1,2,3,4,5,6}.
- Элементарное событие – число выпавших очков
- Примеры событий: A={1,2} – выпало одно или два очка; B={1,3,5} – выпало нечетное число очков.

- **Достоверным** называется событие, которое обязательно происходит в результате эксперимента, т.е. единственное событие, включающее все элементарные исходы
- **Невозможным** называется событие, которое не может произойти в результате эксперимента, т.е. событие не содержащее ни одного элементарного исхода – пустое множество.

1. Основные понятия теории вероятностей.
Вероятность на дискретном пространстве элементарных исходов

- Пространство элементарных событий Ω назовем дискретным, если оно конечно либо счётно. Чтобы определить вероятность любого события A на дискретном пространстве элементарных событий, достаточно присвоить вероятность $p_i \in [0,1]$ каждому элементарному исходу ω_i.

$$\sum_{\omega_i \in \Omega} p_i = 1$$

- Вероятность события A

$$p(A) = \sum_{\omega_i \in A} p_i$$

1. Основные понятия теории вероятностей.
Свойства вероятности на дискретном пространстве элементарных исходов

- $0 \le P(A) \le 1;\ P(\Omega) = 1;\ P(\overline{A}) = 1 - P(A);$
- Если A и B несовместны(наступление события A не зависит от наступления события B и наоборот), то $P(A \cup B) = P(A) + P(B);$
- В общем случае $P(A \cup B) = P(A) + P(B) - P(A \cap B);$
- Если $A \subseteq B$, то $P(A) \le P(B);$
- Когда $A = \varnothing$, $P(A) = 0$.

1. Основные понятия теории вероятностей.
Классическое определение вероятности

- Пусть пространство элементарных событий состоит из конечного числа N элементов: $\Omega=\{\omega_1, \omega_2, \ldots, \omega_N\}$ и все элементарные события **равновозможны.** Тогда вероятность любого из этих событий принимается равной 1/N.
- Если событие A состоит из k элементарных равновозможных исходов $A=\{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}\}$, то **вероятность этого события равна отношению** k/N.

- Пример 1. В урне имеется 10 шаров: 3 белых и 7 синих. Из урны наугад вынимается один шар. Какова вероятность, что этот шар белый? (3/10).

1. Основные понятия теории вероятностей.
Классическое определение вероятности

- Пример 2. В партии из 50 деталей 5 нестандартных. Определить вероятность того, что среди выбранных наудачу для проверки 6-ти деталей 2 окажутся нестандартными.
- Решение.

A={из 6 выбранных наудачу деталей 2 - нестандартные};

Общее число всех возможностей выбора 6 деталей из 50 равно $n = C_{50}^{6}$. Число способов выбрать 2 нестандартные детали из 5 нестандартных, находящихся в партии равно C_5^2. Каждому такому выбору соответствует C_{45}^{4} способов выбора стандартных деталей из 45, имеющихся в партии. По правилу произведения число случаев, благоприятствующих A равно $m = C_5^2 \cdot C_{45}^{4}$. Тогда

$$P(A) = \frac{m}{n} = \frac{C_5^2 \cdot C_{45}^{4}}{C_{50}^{6}} \approx 0{,}08.$$

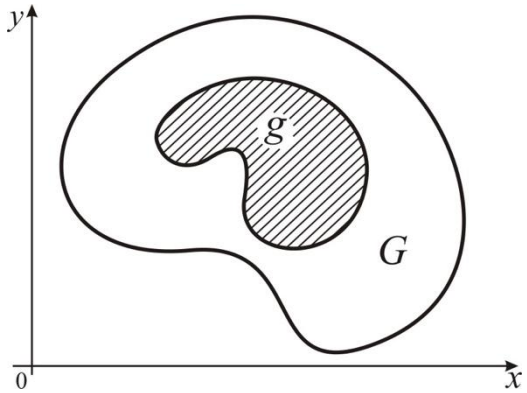
1. Основные понятия теории вероятностей.
Вероятность и частота

- Формула для непосредственного подсчета вероятностей, исходя из классического определения, верна только для **равновозможных** исходов.
- В случае неравновозможных исходов (несимметричная игральная кость и т.п.) вероятность события $p(A)$ необходимо определить по-другому.
- Если произведена серия из n опытов, в каждом из которых событие A произошло m раз, то частость события $p^*(A) = \frac{m}{n}$.
- **Теорема Бернулли**

$$\forall \varepsilon > 0 \lim_{n \to \infty} P(|p^*(A) - p(A)| > \varepsilon) = 0.$$

1. Основные понятия теории вероятностей.
Геометрическое определение вероятности

- Для испытаний с бесконечным числом исходов классическое определение вероятности неприменимо.
- Тогда вводят понятие геометрической вероятности, как вероятности попадания точки в область (отрезок, часть плоскости, часть n-мерного пространства).
- Пример: случайное бросание точки в область G, причем все точки этой области равноправны. Событие A – попадание точки в область g.
- **Геометрической вероятностью** события A называют



$$p(A) = \frac{S_g}{S_G}$$

В общем случае:

$$p(A) = \frac{mes_g}{mes_G}$$

1. Основные понятия теории вероятностей.
Геометрическое определение вероятности

- Пример.

Два студента А и В условились встретиться в определенном месте во время перерыва между 13 ч и 13 ч 50 мин. Пришедший первым ждет другого в течение 10 мин., после чего уходит. Чему равна вероятность их встречи, если приход каждого из них в течение указанных 50 минут может произойти наудачу и моменты прихода независимы?

- Решение.

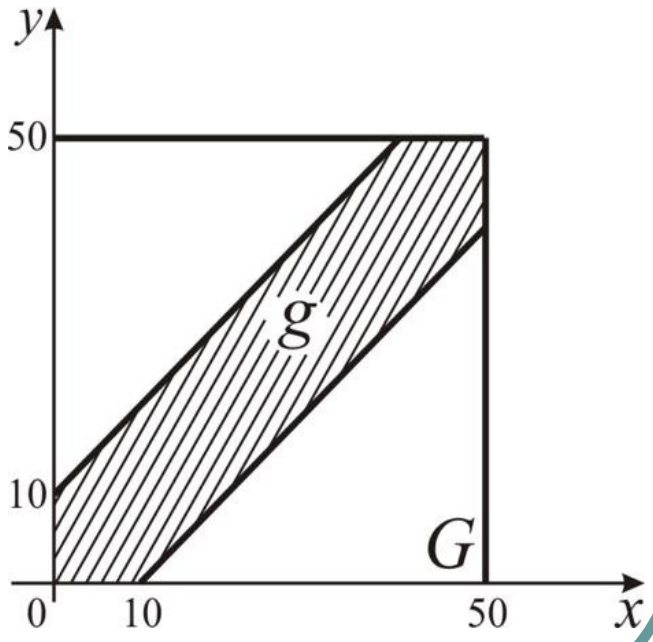
Пусть $0 \le x \le 50$ – время прихода студента А, $0 \le y \le 50$ – время прихода студента В. Исходы изобразим точками квадрата со стороной 50.

$G = \{(x, y): 0 \le x \le 50, 0 \le y \le 50\}$.

Благоприятствующая событию С={студенты А и В встретятся} область

$g = \{|y - x| \le 10: 0 \le x \le 50, 0 \le y \le 50\} \leftrightarrow$

$\leftrightarrow x - 10 \le y \le x + 10.$



$$P(C) = \frac{S_g}{S_G} = \frac{50^2 - 40^2}{50^2} = 0{,}36$$

1. Основные понятия теории вероятностей.
Условная вероятность

- Пример. Игральная кость подбрасывается один раз. Известно, что выпало более трех очков. Какова при этом вероятность, что выпало четное число очков?
- Решение

а) $\Omega = \{4,5,6\}$, $A=\{4,6\}$. $p(A)=2/3$.

б) $\Omega = \{1,2,3,4,5,6\}$; $B = \{4,5,6\}$. Вопрос: какова вероятность того, что при осуществлении B происходит $A =\{4,6\}$: $p(A|B)$?

$$p(A|B) = p(A \cap B)/P(B) =(2/6)/(3/6)=2/3.$$

- **Условной вероятностью** события A по отношению к событию B $p(A|B)$ называют вероятность события A, найденную при условии, что произошло событие B

1. Основные понятия теории вероятностей.
Правило умножения вероятностей событий

- **Правило умножения вероятностей**: Вероятность произведения двух событий равна произведению вероятности одного из этих событий на условную вероятность другого, найденную в предположении, что первое событие произошло, т.е. $p(AB)=p(A)p(B|A)$ или $p(AB)=p(B)p(A|B)$
- События A и B называются **независимыми**, если $p(A|B) = p(A)$ и $p(B|A) = p(B)$. Для независимых событий $p(AB)=p(A)p(B)$.

- Пример. В первом ящике 2 белых и 10 красных шаров, во втором ящике – 8 белых и 4 красных. Из каждого ящика вынули по шару. Какова вероятность, что оба шара белые?
- Решение. A={появление белого шара из первого ящика}, B={появление белого шара из второго ящика}. A и B – независимы. $p(AB)=p(A)p(B)=2/12 \cdot 8/12 = 1/9$

1. Основные понятия теории вероятностей.
Формула полной вероятности

- Вероятность p(A) появления события A, которое может произойти только совместно с одним из событий B_1, B_2, …. B_n , образующих полную группу попарно несовместных событий:

$$\forall i\ p(B_i) > 0;\ \forall i \neq j\ B_i \cap B_j = \emptyset;\ \sum_{i=1}^{n} B_i = \Omega$$

вычисляется по формуле полной вероятности:

$$p(A) = \sum_{i=1}^{n} p(B_i)p(A|B_i), \text{причем } \sum_{i=1}^{n} p(B_i) = 1.$$

- События B_i обычно называют гипотезами, а числа $p(B_i)$ – вероятностями гипотез.

1. Основные понятия теории вероятностей.
Формула полной вероятности

- Пример. Имеется четыре одинаковых ящика с электрическими лампочками, причем первый ящик содержит 10 исправных и 2 бракованные лампочки, второй и третий ящики содержат по 5 исправных и по 5 бракованных лампочек, а четвертый ящик содержит только 10 исправных лампочек. Наудачу выбирается один ящик и из него одна лампочка. Какова вероятность того, что эта лампочка окажется исправной?
- Решение. Событие A={выбор исправной лампочки}. Гипотезы B_i={выбор i-го ящика}. События B_i образуют полную группу событий, $p(B_i)=1/4$. $p(A|B_1)=10/12=5/6$; $p(A|B_2)= p(A|B_3)=5/10=1/2$; $p(A|B_4)=10/10=1$. Тогда по формуле полной вероятности $p(A)=p(B_1)p(A|B_1)+ p(B_2)p(A|B_2)+ p(B_3)p(A|B_3)+ p(B_4)p(A|B_4) = 1/4 \cdot 5/6+ 1/4 \cdot 1/2+ 1/4 \cdot 1/2+ 1/4 \cdot 1=17/24$

1. Основные понятия теории вероятностей.
Формула Байеса

- Условная вероятность гипотезы B_i в предположении, что событие A уже имеет место, определяется по **формуле Байеса**:

$$p(B_i|A) = \frac{p(AB_i)}{p(A)} = \frac{p(B_i)p(A|B_i)}{\sum_{i=1}^{n} p(B_i)p(A|B_i)}$$

- Пример. Три организации представили в контрольное управление счета для выборочной проверки: первая 15 счетов, вторая – 10, третья – 25. Вероятности правильного оформления счетов у этих организаций соответственно таковы: 0,9; 0,8; 0,85. Был выбран один счет, и он оказался правильным. Определить вероятность того, что этот счет принадлежит второй организации.
- Решение. Событие A={представлен правильно оформленный счет}, гипотезы B_i={правильно оформленный счет представила i-я организация}. События B_i образуют полную группу несовместных событий, при этом $p(B_1)=15/50=0,3$, $p(B_2)=10/50=0,2$, $p(B_3)=25/50=0,5$. По условию $p(A|B_1)=0,9$; $p(A|B_2)=0,8$; $p(A|B_3)=0,85$. По формуле полной вероятности $p(A)=0,3\cdot0,9+0,2\cdot0,8+0,5\cdot0,85=0,855$.
По формуле Байеса $p(B_2|A)=p(B_2)p(A|B_2)/p(A)=0,2\cdot0,8/0,855 \approx 0,19$

2. Случайные величины и их характеристики
Понятие случайной величины

- **Случайной величиной** называется величина которая в результате опыта принимает то или иное числовое значение, причем заранее, до опыта, неизвестно, какое именно.

- Современная теория вероятностей предпочитает оперировать не с событиями, а с соответствующими им случайными величинами.

- **Дискретные** случайные величины принимают конечное или счетное множество значений. Примеры: число попаданий в цель при трех выстрелах, число вызовов, поступавших на телефонную станцию за сутки.
- Случайные величины, значения которых непрерывно заполняют некоторый промежуток (конечный или бесконечный) числовой оси называют **непрерывными**. Примеры: скорость космического аппарата при выходе на орбиту, ошибка взвешивания тела на аналитических весах.

2. Случайные величины и их характеристики
Закон распределения

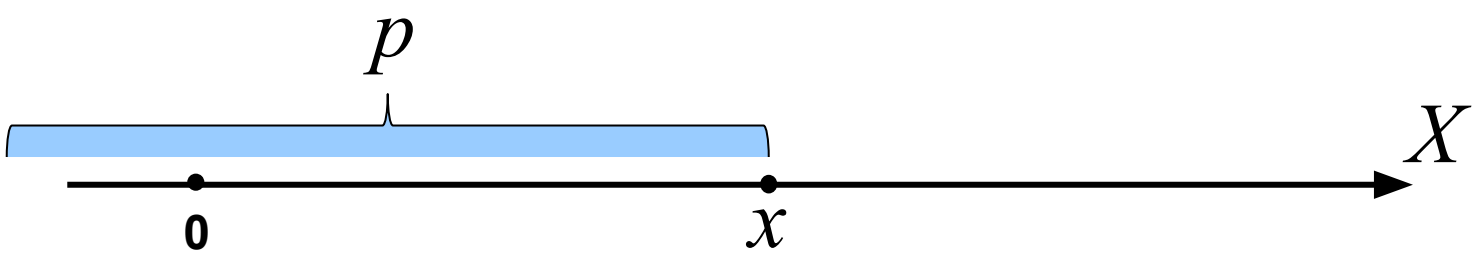
- Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.
- Закон распределения может быть задан аналитически, графически, для дискретной случайной величины – в виде таблицы:

X:	x_1	x_2	...	x_n	...
	p_1	p_2	...	p_n	...

2. Случайные величины и их характеристики
Функции распределения случайных величин

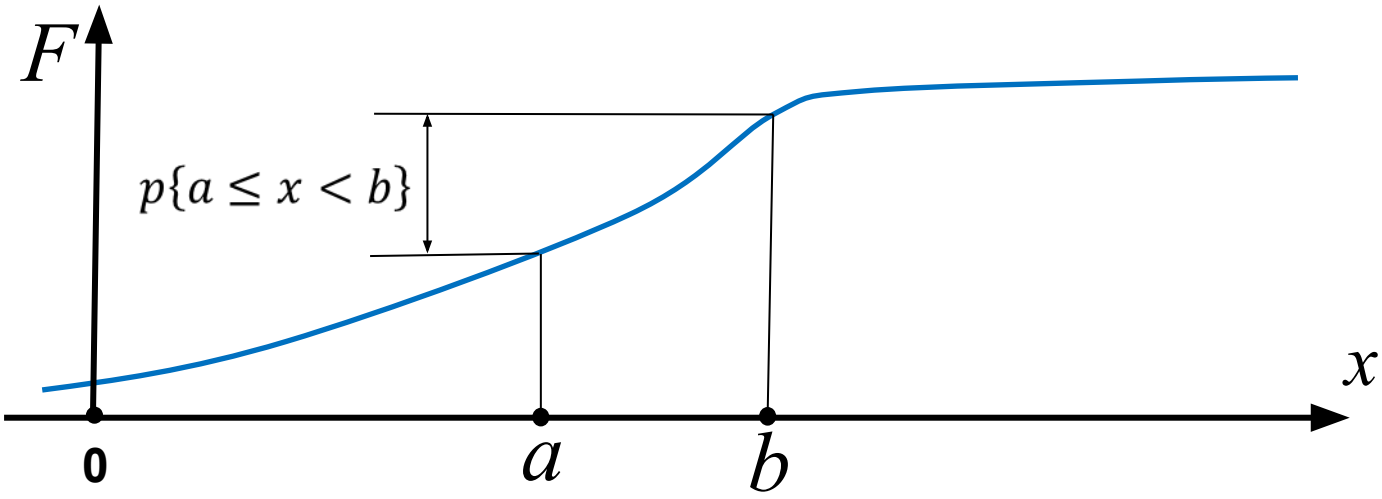
- Функцией распределения (или интегральной функцией распределения) случайной величины X называется функция F(x), которая для любого числа x∈ R равна вероятности события, состоящего в том, что случайная величина X примет значение, меньшее чем заданное x (аргумент функции).

$$F(x) = p\{X < x\}$$



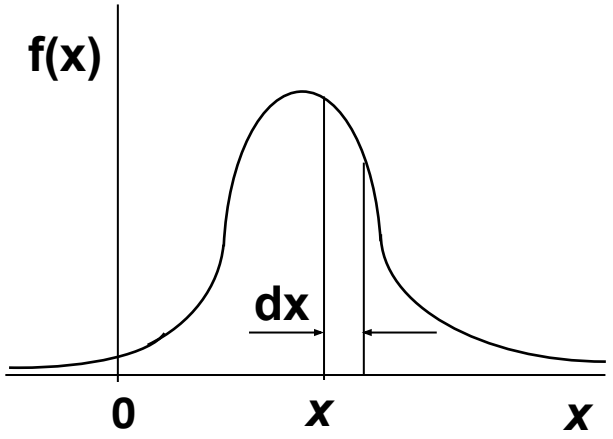
2. Случайные величины и их характеристики
Свойства функции распределения

- $F(x)$ – неубывающая функция своего аргумента, т.е. если $x_2 > x_1$, то $F(x_2) > F(x_1)$
- $F(-\infty) = 0; F(+\infty) = 1$
- $0 \le F(x) \le 1$
- $p\{a \le x < b\} = F(b) - F(a)$
- $F(x)$ – непрерывна слева в любой точке: $F(x-0) = F(x), x \in R$



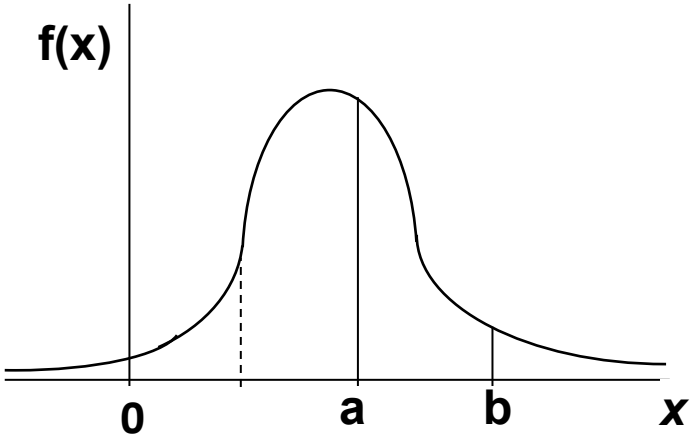
2. Случайные величины и их характеристики
Плотность распределения непрерывной случайной величины

- Плотностью распределения непрерывной случайной величины называется функция $f(x) = \frac{dF}{dx}$.

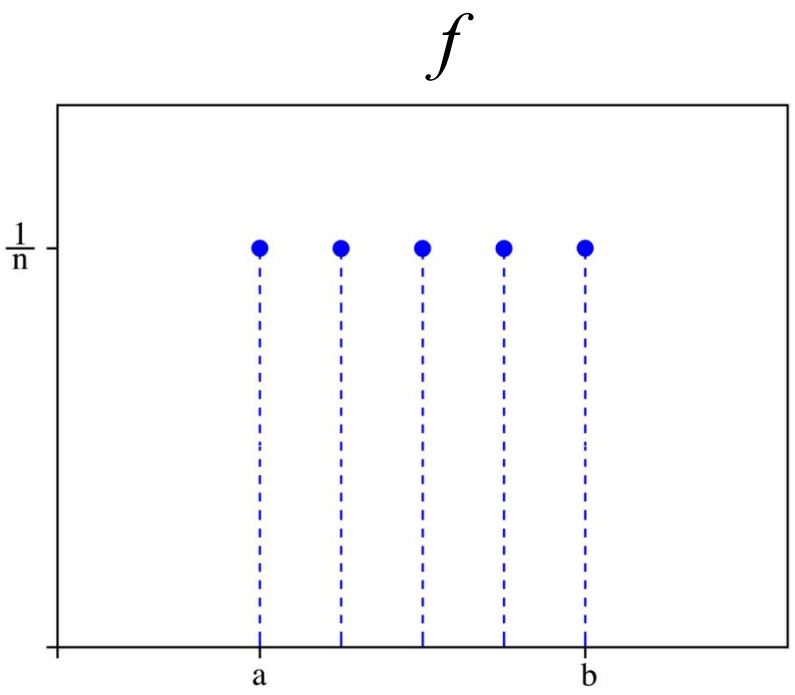


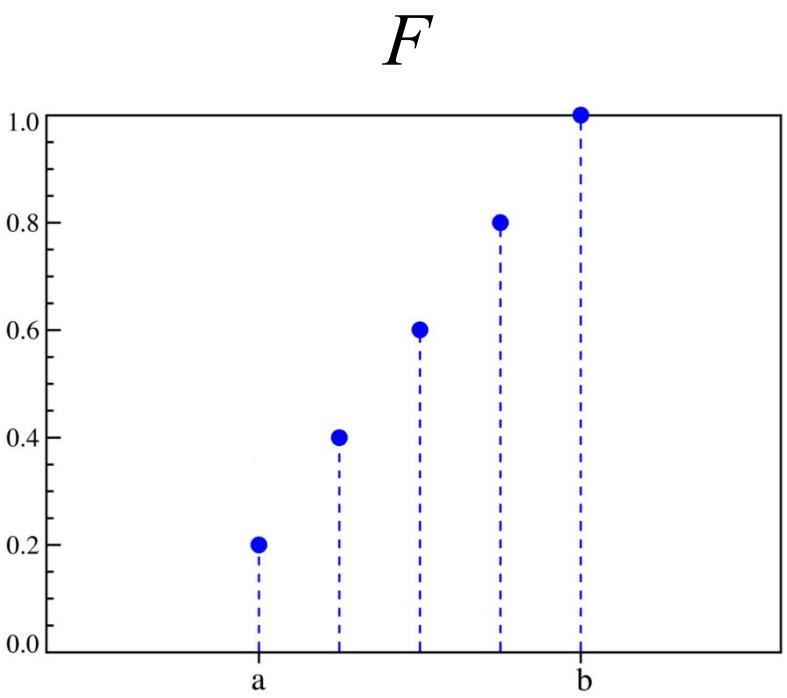
- $p\{x \le X < x + dx\} = f(x)dx$

- $p\{a < X < b\} = \int_a^b f(x)dx$



2. Случайные величины и их характеристики
Дискретное равномерное распределение

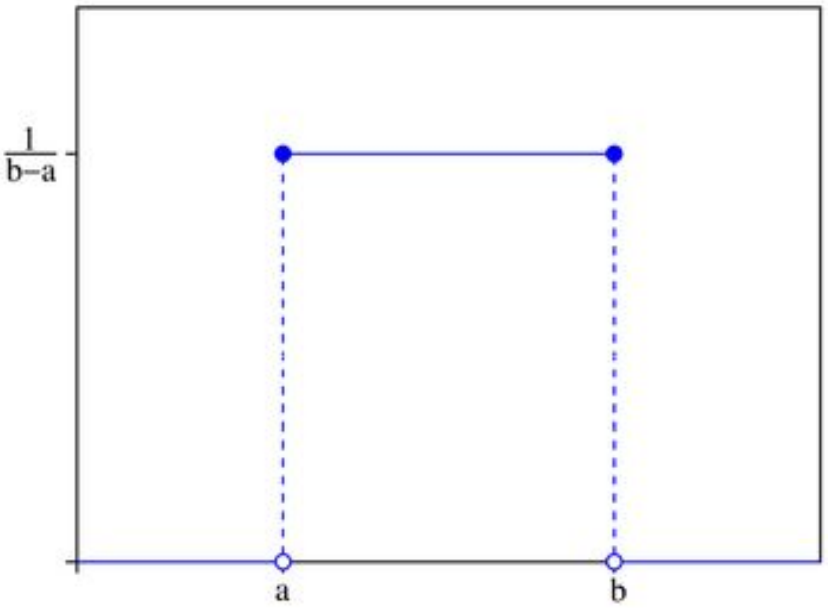


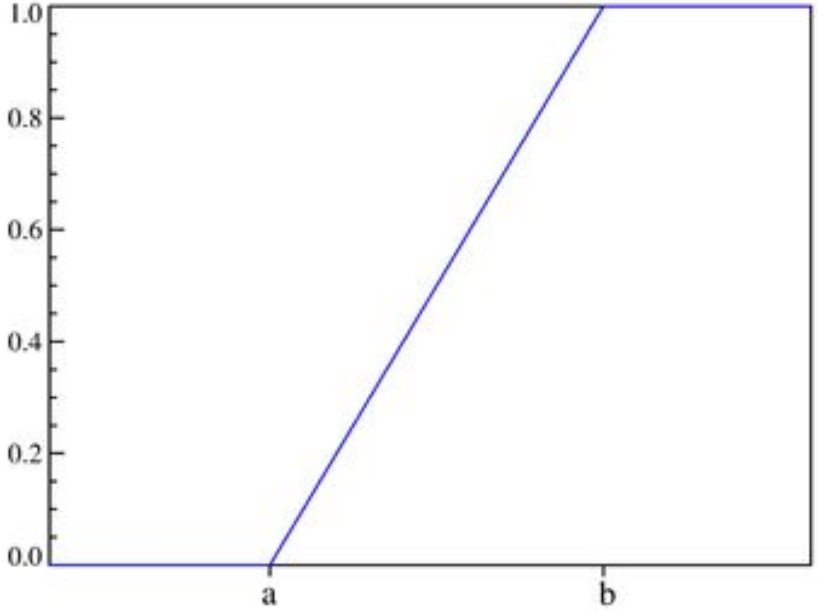


2. Случайные величины и их характеристики
Непрерывное равномерное распределение

$$f(x) = \begin{cases} \dfrac{1}{b-a}, x \in [a,b] \\ 0, x \notin [a,b] \end{cases}$$

$$F(x) = \begin{cases} 0, x < a \\ \dfrac{x-a}{b-a}, a \le x < b \\ 1, x \ge b \end{cases}$$



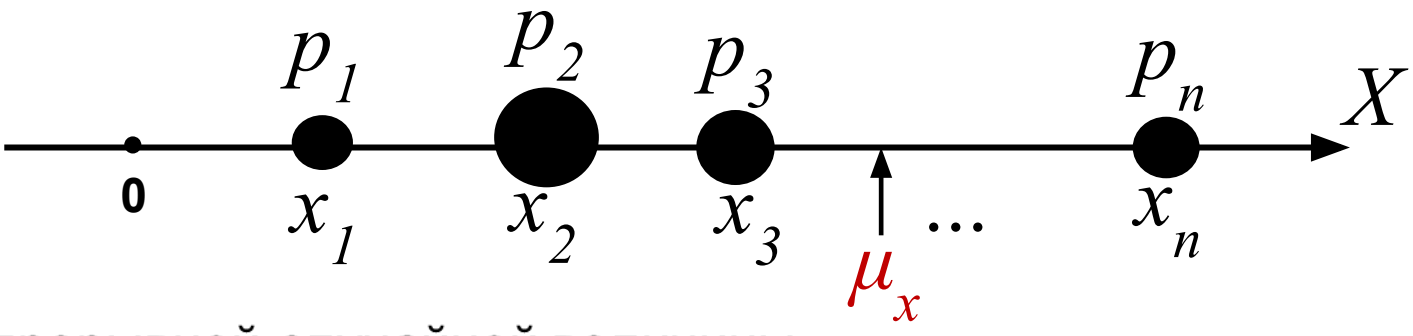


Основные характеристики случайных величин

- Математическое ожидание (среднее значение):

$$M[X] = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i}; \qquad \sum_{i=1}^{n} p_i = 1;$$

$$M[X] = \sum_{i=1}^{n} x_i p_i = \mu_x;$$



Для непрерывной случайной величины

$$M[X] = \mu_x = \int_{+\infty} x f(x) dx$$

2. Случайные величины и их характеристики
Основные характеристики случайных величин

- Дисперсия

$$D_x = D[X] = \sum_{i=1}^{n} (x_i - \mu_x)^2 p_i ;$$

$$D_x = D[X] = \int_{-\infty}^{+\infty} (x - \mu_x)^2 f(x) dx$$

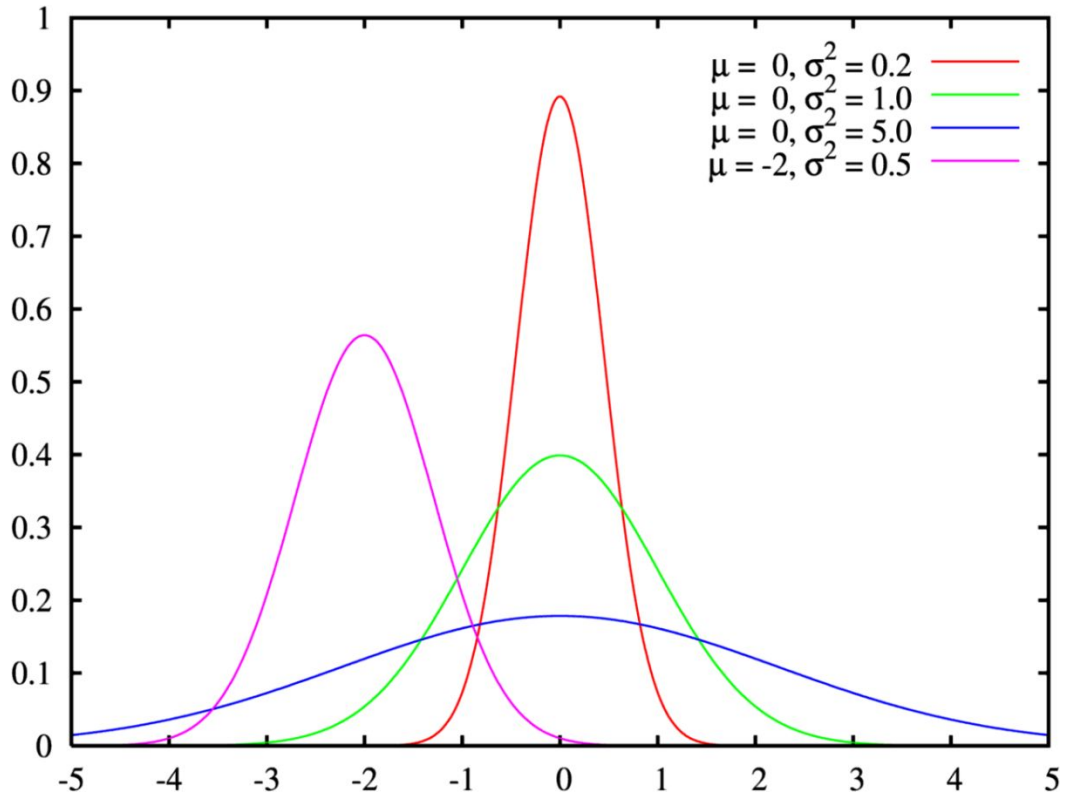
$$D_x = D[X] = M[X^2] - M[X]^2$$

- Среднее квадратичное отклонение

$$\sigma[X] = \sigma_x = \sqrt{D[X]} = \sqrt{D_x}$$

2. Случайные величины и их характеристики
Нормальное распределение

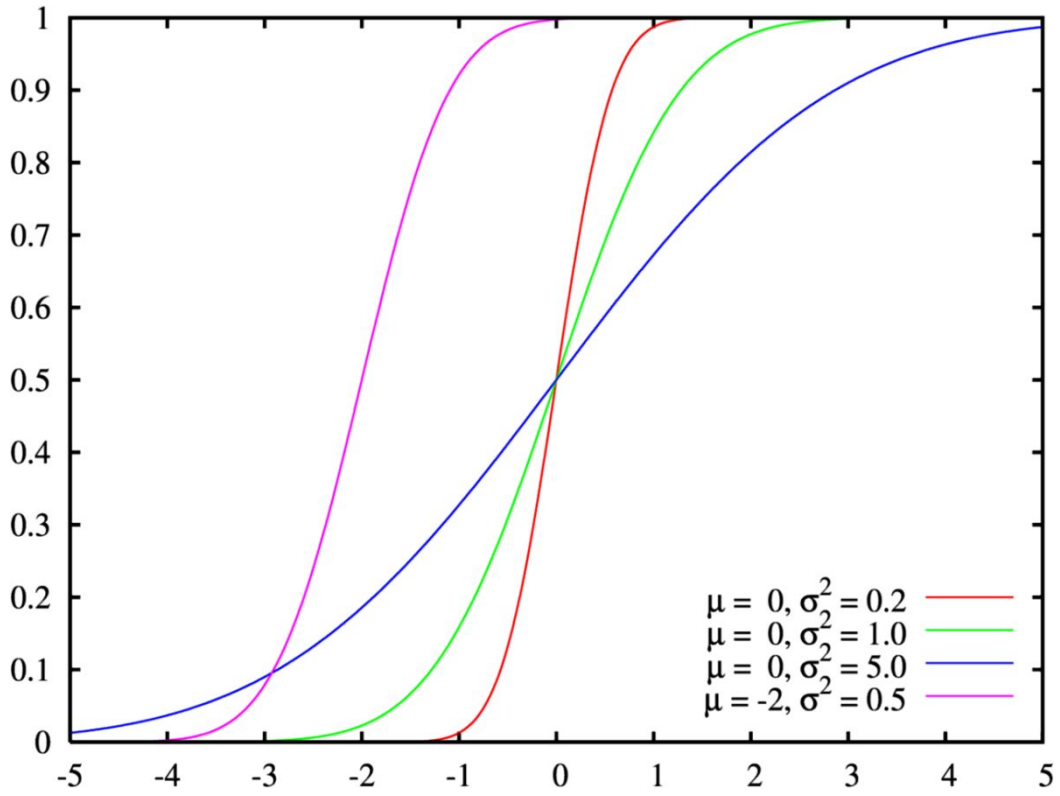
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



2. Случайные величины и их характеристики
Нормальное распределение

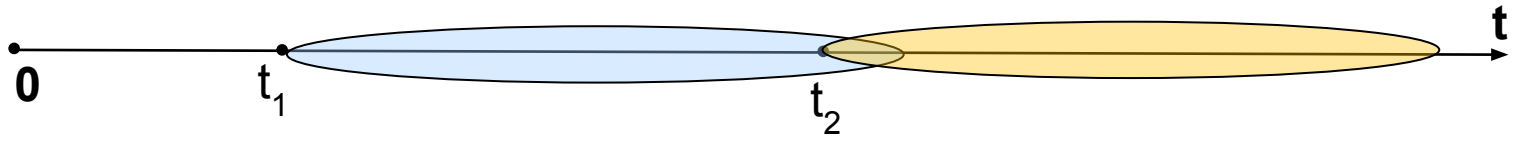
$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right], \qquad \operatorname{erf} x = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt$$

- функция Лапласа или интеграл вероятности



2. Случайные величины и их характеристики
Понятие случайного процесса

- Случайный процесс – семейство случайных величин $\{X(t), t \in T\}$, заданных на вероятностном пространстве $(\Omega, \mathcal{F}, P)$ и некотором промежутке T.
- Если множество T дискретно, то $X(t)$ образуют случайную последовательность.
- **Стационарный процесс** – все конечномерные характеристики последовательности инвариантны относительно сдвигов по времени.



- Если при определении характеристик процесса усреднение по статистическому ансамблю можно заменить усреднением по времени, процесс называется **эргодическим**.
- Случайный процесс $\{X(t), t \in T\}$ называется **Марковским**, если величины $X(t)$ в любой момент времени $t \in T$ принимают значения в конечном или счетном множестве S и состояние, которое примет система в момент $t \in T$ при заданных состояниях в предшествующие моменты времени, зависит

2. Случайные величины и их характеристики
Основные задачи статистики

- Предмет математической статистики – разработка методов регистрации, описания и анализа статистических экспериментальных данных, получаемых в результате наблюдения массовых случайных явлений.
- Основные задачи математической статистики:
  1. Задача определения закона распределения случайной величины (или системы случайных величин) по статистическим данным
  2. Задача проверки правдоподобия гипотез
  3. Задача нахождения неизвестных параметров распределения

«There are three kinds of lies: lies, damned lies, and statistics.»
Приписывается премьер-министру Великобритании Бенджамину Дизраэли.

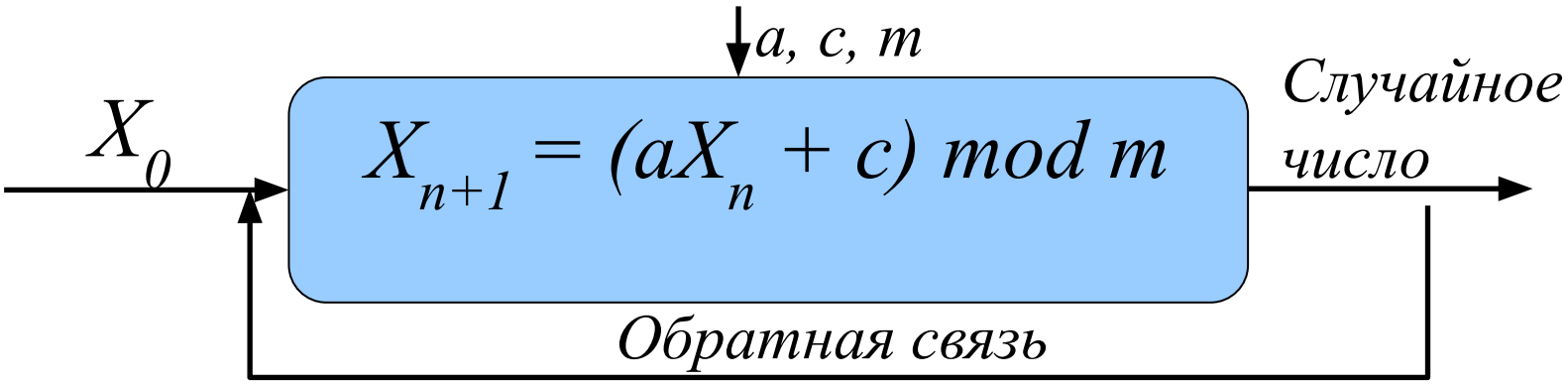
2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

- Источники настоящих последовательностей случайных чисел – случайные природные процессы: оптические квантовые эффекты (отражение фотонов от полупрозрачного зеркала), радиоактивный распад, дробовой шум в радиоэлектронных приборах за счет дискретности носителей тока, детектирование космического излучения и т.п.).

- Компьютер – детерминированная система. С его помощью можно генерировать только псевдослучайные последовательности.

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

- Линейный конгруэнтный (рекурсивный) метод (Lehmer, 1949):



$m > 0$, $0 < a \le m$, $0 \le c \le m$, начальное значение X_0: $0 < X_0 \le m$.

- Модуль m должен быть достаточно большим, т.к. период не больше m. Удобно связать m с длиной слова компьютера и использовать $m=2^e - 1$, либо $m=2^e + 1$ для e-разрядной машины, а еще лучше – m наибольшее простое, меньшее 2^e.
- Длина периода равна m в следующем случае: c и m – взаимно простые числа, $b = a - 1$ кратно p для любого p, являющегося множителем m, b кратно 4, если m кратно 4.

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

- Генератор MS FORTRAN: $m = 2^{31}-1,\ c=0,\ a=48271$

$$X_{n+1} = 48271X_n \ mod\ (2^{31}-1)$$

- Генератор Парка-Миллера: $m = 2^{31}-1,\ c=0,\ a=75$

$$X_{n+1} = 75X_n \ mod\ (2^{31}-1)$$

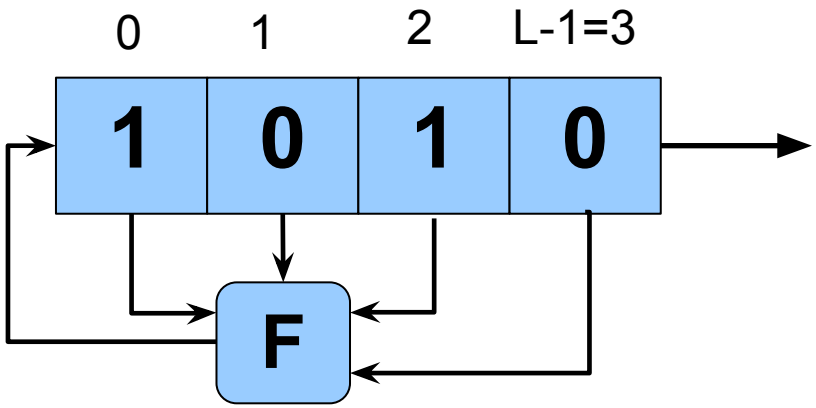
- Нелинейные генераторы:

$$X_{n+1} = (aX_n^3 + bX_n^2 + cX_n + d) mod\ m$$

- Суперпозиция нескольких конгруэнтных генераторов посредством нелинейной функции.

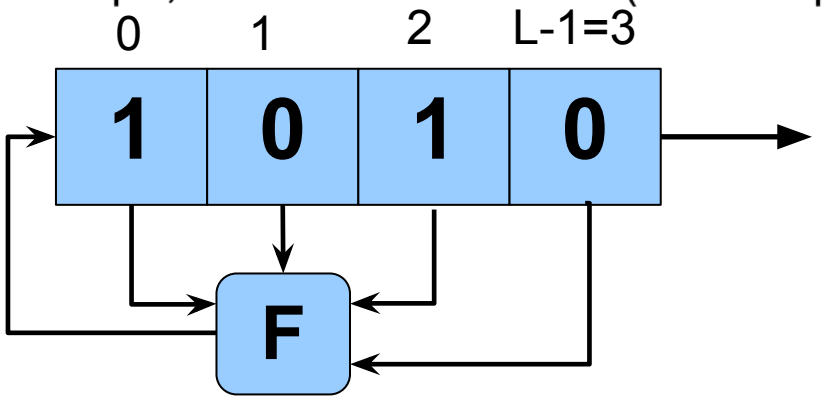
2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

- Линейные регистры с обратной связью



2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

- Алгоритм работы линейного регистра с обратной связью:
  1. Прочитать бит из ячейки L-1 и направить его в выходную последовательность
  2. Вычислить новое значение для ячейки 0 с помощью заданной функции обратной связи F, используя текущие значения ячеек
  3. Выполнить сдвиг вправо (содержимое каждой i-й ячейки перемещается в ячейку i+1)
  4. Записать в ячейку 0 бит, ранее вычисленный функцией обратной связи F
- Чаще всего, $F = (c_0 s_0) \oplus (c_1 s_1) \ldots \oplus (c_{L-1} s_{L-1});\ c_i = \{0,1\}$
- Периодичность: $T \leq 2^L - 1$. Для того, чтобы период был максимальным, необходимо (но не достаточно), чтобы число отводов было четным, их номера, взятые все вместе (не попарно) были взаимно простыми.



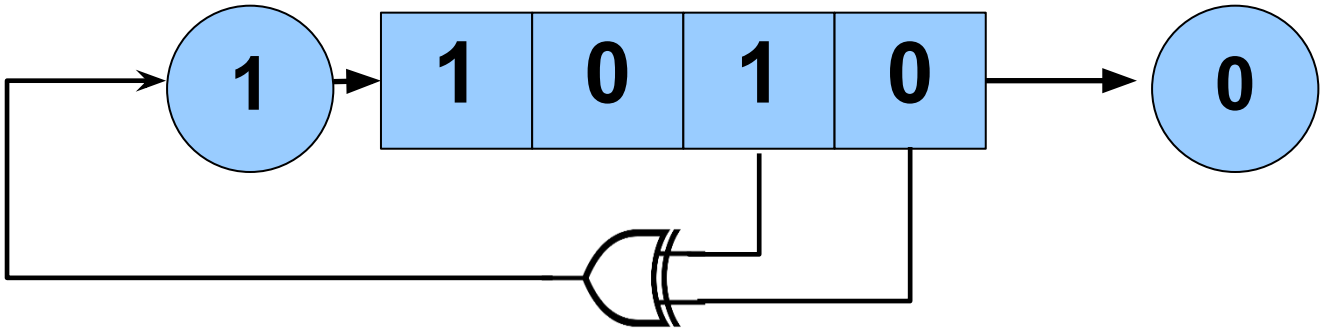
2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

- $\oplus$ XOR - Исключающее ИЛИ (сложение по модулю 2);

a	b	
0	0	0
1	1	0
0	1	1
1	0	1

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

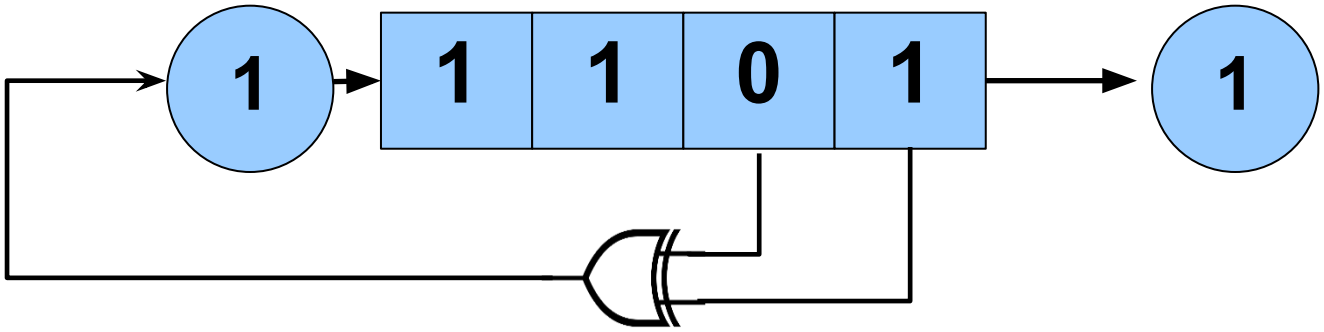
- Линейные регистры с обратной связью



Выходная последовательность:
0

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

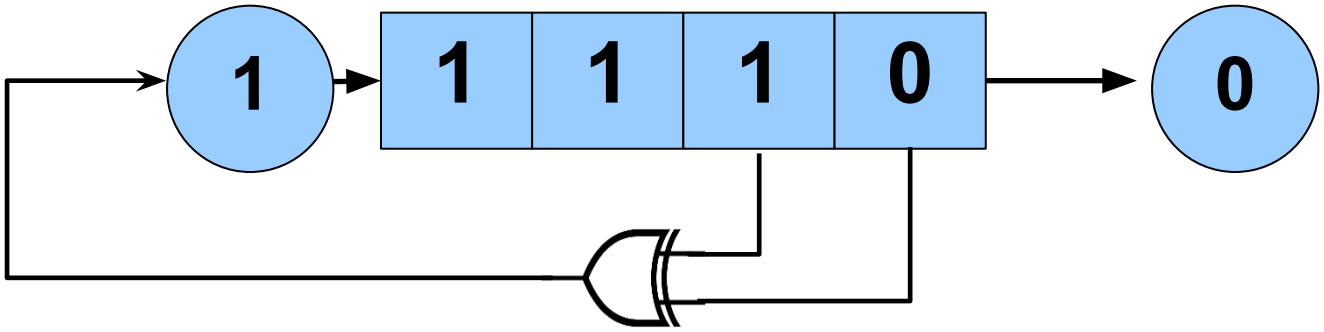
- Линейные регистры с обратной связью



Выходная последовательность:
01

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

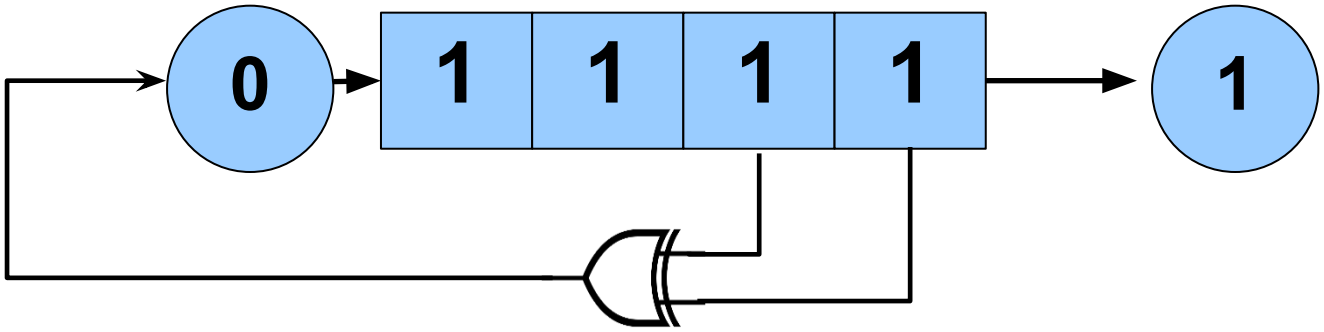
- Линейные регистры с обратной связью



Выходная последовательность:
010

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

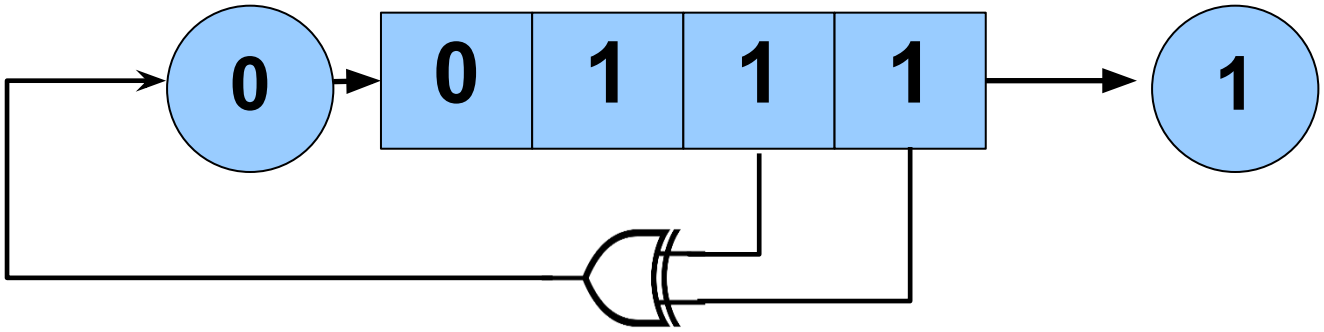
- Линейные регистры с обратной связью



Выходная последовательность:
0101

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

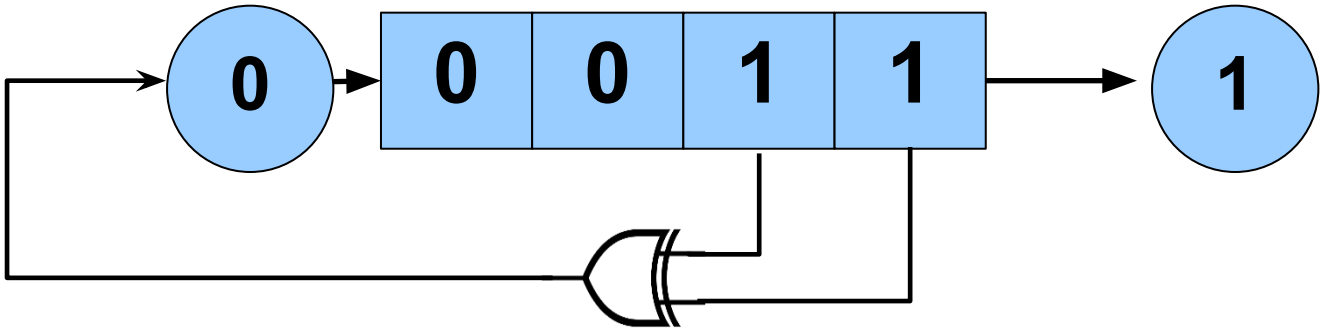
- Линейные регистры с обратной связью



Выходная последовательность:
0101 1

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

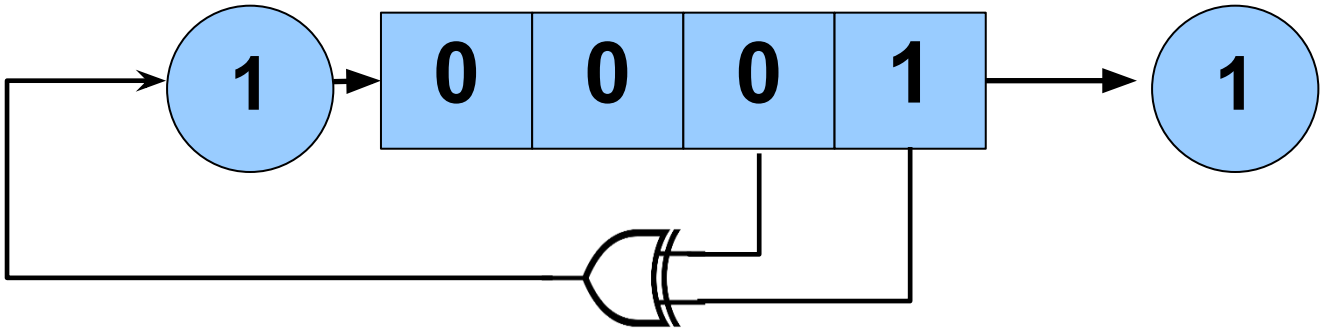
- Линейные регистры с обратной связью



Выходная последовательность:
0101 11

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

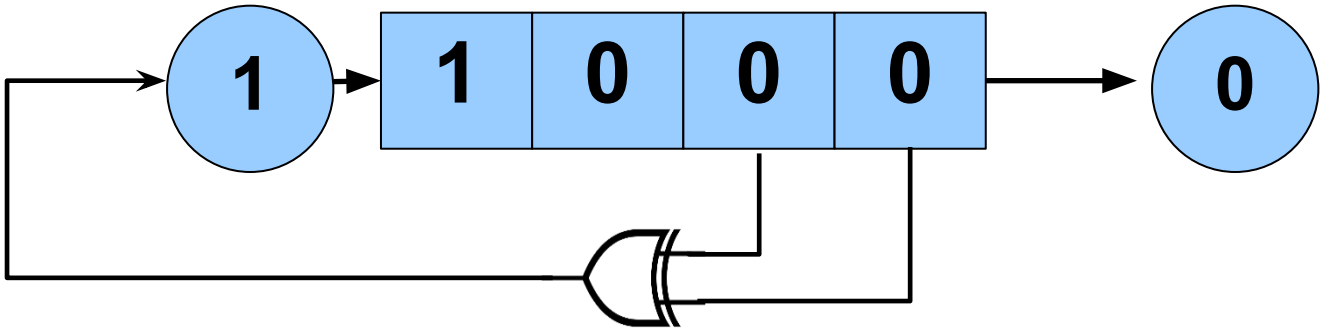
- Линейные регистры с обратной связью



Выходная последовательность:
0101 111

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

- Линейные регистры с обратной связью



Выходная последовательность:
0101 1110 …

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

Недостатки генераторов псевдослучайных чисел:

- Конечный период
- Последовательные значения не являются независимыми.
- Некоторые биты «менее случайны», чем другие.
- Неравномерное одномерное распределение.
- Обратимость.

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

Основные критерии криптостойкости:

- Нет аналитической зависимости между последовательно сгенерированными числами
- Зная предыдущие числа, нельзя найти следующее (атака из прошлого)
- Зная последующие числа, нельзя восстановить предшествующие (атака из будущего)
- Вероятность появления любого числа в последовательности одинакова

2. Случайные величины и их характеристики
Генерация псевдослучайных последовательностей

Примеры тестов генераторов псевдослучайных последовательностей

- Частотный тест (равновероятность 0 и 1 в последовательности)
- Блочный тест на частоту (последовательность разбивается на блоки длиной M бит и для каждого рассчитывается, насколько вероятность появления 1 близка к ½)
- Тест распределения на плоскости. Последовательность чисел группируется парами, которые рассматриваются как координаты на двумерном графике. Отображение этих точек на плоскости является результатом теста. Для случайной последовательности расположение точек на плоскости будет хаотичным, а при росте выборки плоскость полностью будет заполнена точками. Признаком неслучайной последовательности является наличие на полученном изображении «узоров» (явно выраженных вертикальных либо горизонтальных линий, периодических рисунков и т.д.).