Имитация случайности и вероятностные законы

Simulation (англ.) — моделирование

Язык программирования R

R — язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом в рамках проекта GNU. Язык создавался как аналогичный языку S, разработанному в Bell Labs и является его альтернативной реализацией, хотя между языками есть существенные отличия, но в большинстве своём код на языке S работает в среде R. Изначально R был разработан сотрудниками статистического факультета Оклендского университета Россом Айхэкой (англ. *Ross Ihaka*) и Робертом Джентлменом (англ. *Robert Gentleman*) (первая буква их имён — R).

Сайт: r-project.org

R широко используется как статистическое программное обеспечение для анализа данных и фактически стал стандартом для статистических программ.

В R используется интерфейс командной строки, хотя доступны и несколько графических интерфейсов пользователя: R Commander, RKWard, RStudio.

Сайт: rstudio.com

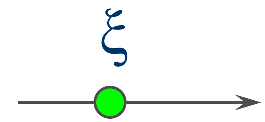
Некоторые простые команды языка R

- Определить переменную x и присвоить ей значение 7: x=7
- Посмотреть значения объекта: набрать имя, нажать Enter
- Создать вектор v и заполнить его числами от 1 до 5: v=1:5
- Создать вектор v и задать значения компонент: v=c(2, 5, 4)
- Упорядочить значения вектора v по возрастанию: u=sort(v)
- Узнать, какие аргументы имеет функция: ?имя функции
- Найти сумму всех компонент вектора v: sum(v)
- Преобразовать компоненты v по условию: w=ifelse(v>3,1,-1)
- Создать матрицу m размерности 3 x 2 и заполнить её числами 1, 7, 3, 5, 4, 6 по строкам:
  m=matrix(c(1,7,3,5,4,6), nrow=3, ncol=2, byrow=TRUE)
- Выбрать элемент с индексами (1, 2) из матрицы: z=m[1,2]
- Выбрать 1-ю строку матрицы (таблицы данных): x=m[1,]
- Выбрать 2-й столбец матрицы (таблицы данных): y=m[,2]
- Удалить строку 3 из матрицы (таблицы данных): r=m[-3,]
- Удалить строки 2 и 3 из матрицы: r=m[-c(2,3),]
- Выбрать из таблицы d строки по условию и заданные столбцы:
  s=subset(d, Sex=="female" & Age<25, select=c("Name", "Tel"))
- Записать таблицу данных d в папку c:/my_dir в файл d_file.txt:
  write.table(d, file="c:/my_dir/d_file.txt)

Основные понятия теории вероятностей

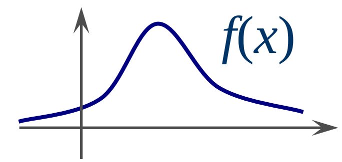
!!!

- Случайная величина



- Функция распределения

- Плотность распределения



- Математическое ожидание и дисперсия $\mathbf{M}\xi,\ \mathbf{D}\xi$

- Независимость случайных величин. вел.

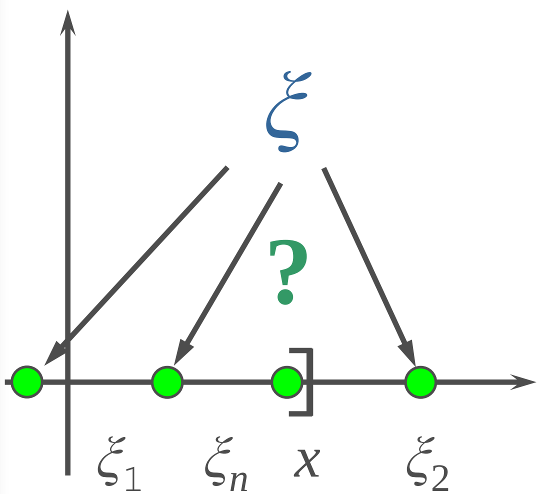
$$\mathbf{P}(\xi \le x,\ \eta \le y) = \mathbf{P}(\xi \le x)\,\mathbf{P}(\eta \le y)$$

- Ковариация и коэффициент корреляции

$$\mathbf{cov}(\xi,\eta) = \mathbf{M}\xi\eta - \mathbf{M}\xi\,\mathbf{M}\eta \qquad \boldsymbol{\rho}(\xi,\eta) = \mathbf{cov}(\xi,\eta)\,/\,\sqrt{\mathbf{D}\xi\,\mathbf{D}\eta}$$

Случайные величины

Представим, что проводится эксперимент, результат которого — действительное число ξ — зависит от случая. Как описать случайную величину ξ, т. е. как сформулировать вероятностный закон её поведения?

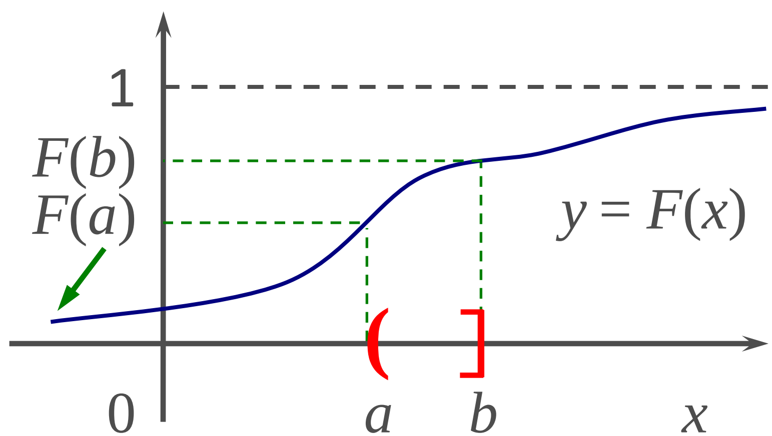


Допустим, что возможно повторить эксперимент несколько раз. Обозначим полученные значения через $\xi_1, \xi_2, \ldots, \xi_n$. Тогда для заданной точки x на прямой можно подсчитать ν_n — количество значений ξ_i, попавших левее x.

Предположим, что существует предел частоты ν_n / n при стремлении n к бесконечности. Этот предел будем называть вероятностью того, что $\xi \leq x$, и обозначать через $\mathbf{P}(\xi \leq x)$.

Функция распределения

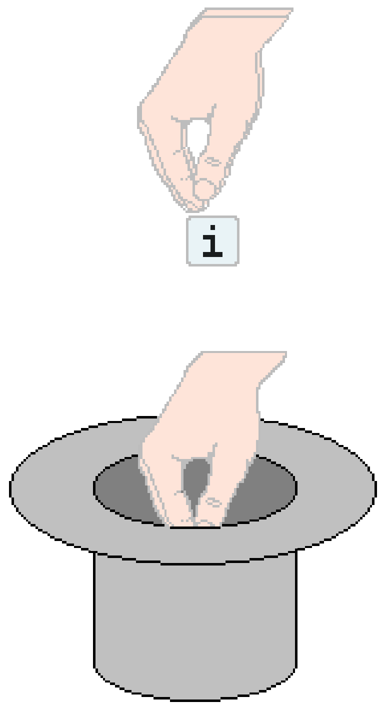
Функция $F(x) = \mathbf{P}(\xi \leq x)$ называется функцией распределения случайной величины ξ. Понятно, что $F(x)$ — неубывающая функция, которая стремится к 0 при $x \to -\infty$ и стремится к 1 при $x \to +\infty$.

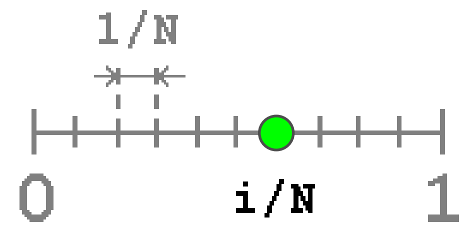


С помощью $F(x)$ можно найти вероятность попадания случайной величины ξ в любой промежуток $(a, b]$ на прямой:

$$\boxed{\mathbf{P}(a < \xi \leq b) = F(b) - F(a).}$$

Выбор точки наудачу из [0, 1]



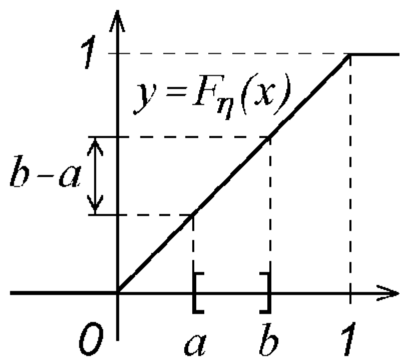


Можно представлять себе, что в шляпе лежат бумажки с номерами от **1** до **N**. Случайно извлекается одна бумажка. Если на ней написан номер **i**, то на отрезок **[0, 1]** ставится точка с координатой **i/N**.

Устремляя **N** к бесконечности, приходим к выбору точки наудачу из отрезка **[0, 1]**. Координату η такой точки называют равномерно распределённой на отрезке **[0, 1]**.

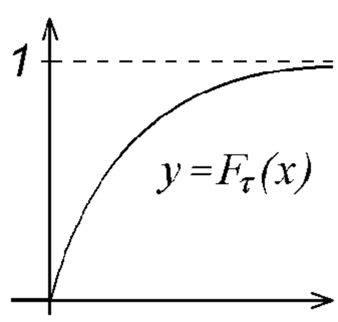
Примеры распределений

Равномерно распределенная случайная величина η



$$F_\eta(x) = \begin{cases} 0 & \text{при } x \leqslant 0, \\ x & \text{при } 0 < x < 1, \\ 1 & \text{при } x \geqslant 1. \end{cases}$$

Показательная (экспоненциальная) случайная величина τ с параметром $\lambda > 0$



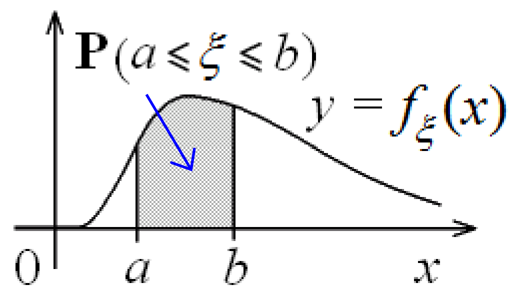
$$F_\tau(x) = \begin{cases} 0 & \text{при } x \leqslant 0, \\ 1 - e^{-\lambda x} & \text{при } x > 0. \end{cases}$$

Плотность случайной величины

Если существует такая неотрицательная функция $f_\xi(x)$, что для любых чисел $a < b$

$$\mathbf{P}(a \leqslant \xi \leqslant b) = \int_a^b f_\xi(x)\,dx,$$

то говорят, что случайная величина ξ имеет плотность $f_\xi(x)$.



Когда плотность существует, её можно найти дифференцированием функции распределения: $f_\xi(x) = F'_\xi(x)$.

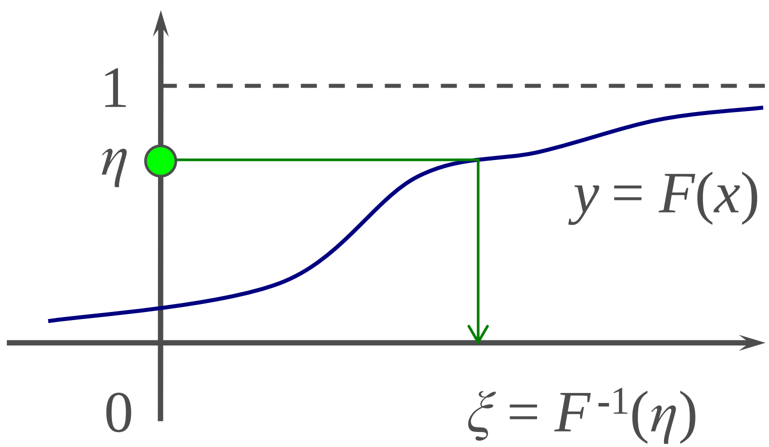
Обратно, положив в верхней формуле $a = -\infty$ и $b = x$, получим, что

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(y)\,dy.$$

Метод обратной функции

Допустим, что функция распределения $F(x)$ непрерывна и строго возрастает. Тогда существует обратная функция $F^{-1}(y)$, которая также строго возрастает, и справедливо следующее

Утверждение. Если случайная величина η равномерно распределена на **[**0, 1**]**, то случайная величина $\xi = F^{-1}(\eta)$ имеет функцию распределения $F(x)$.



Метод обратной функции позволяет моделировать выборку с заданным распределением с помощью датчика случайных чисел.

Доказательство. Так как $0 \le F(x) \le 1$ и $F^{-1}(x)$ возрастает, то

$$F(x) = \mathbf{P}(\eta \le F(x)) = \mathbf{P}(F^{-1}(\eta) \le F^{-1}(F(x))) = \mathbf{P}(\xi \le x).$$

Практическое задание 1

1) Моделируйте в RStudio выборку (вектор u) из 100 равномерно распределённых на отрезке [0, 1] случайных чисел с помощью функции runif (название функции происходит от английских слов random и uniform — случайные и равномерные)

2) Получите формулу для обратной функции к функции распределения показательной случайной величины τ:

$$F_\tau(x) = \begin{cases} 0 & \text{при } x \leqslant 0, \\ 1 - e^{-\lambda x} & \text{при } x > 0 \end{cases}$$

с параметром $\lambda = 5$ (вместо $F_\tau(x)$ запишите переменную y, затем выразите переменную x через переменную y)

3) Применив метод обратной функции, получите выборку t из показательного распределения с параметром $\lambda = 5$

4) В полученной показательной выборке подсчитайте количество значений, оказавшихся больше, чем $3/\lambda = 0{,}6$ (используйте функции ifelse и sum)

5) Вычислите вероятность $\mathbf{P}(\tau > 3/\lambda)$ с помощью функции exp

Важнейшие предельные теоремы теории вероятностей

- **Закон больших чисел**

  При увеличении размера выборки выборочные средние сходятся по вероятности к математическому ожиданию элементов выборки.

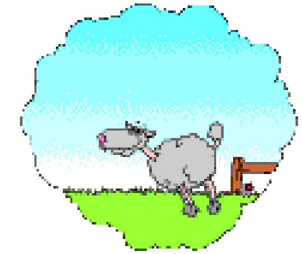
- **Центральная предельная теорема**

  Суммы независимых случайных величин после центрирования и нормирования сходятся к стандартному нормальному закону.

- **Теорема Пуассона (закон редких событий)**

  Если $n \to \infty$, $p \to 0$, $np \to \lambda > 0$,
  то биномиальное распределение приближается к закону Пуассона с параметром λ.

Закон больших чисел

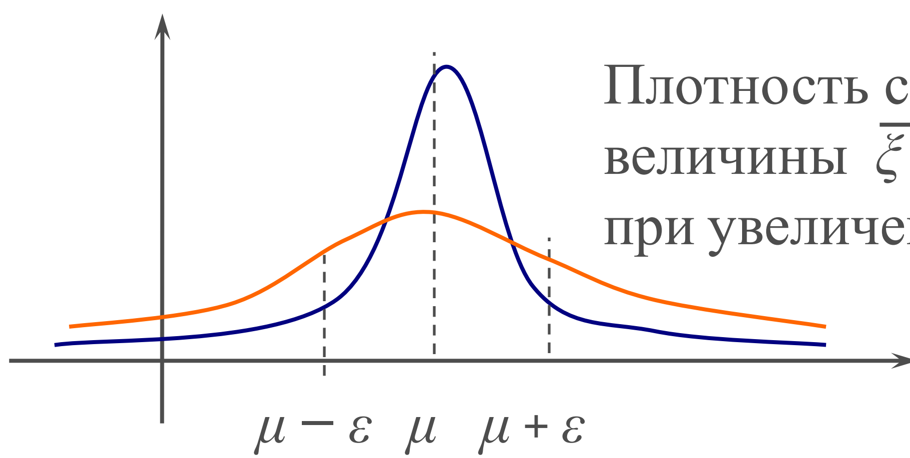


Пусть $\xi_1, \xi_2, \ldots$ — независимые и одинаково распределенные случайные величины с математическим ожиданием $\mu = \mathbf{M}\xi_1$. Рассмотрим $\bar{\xi} = (\xi_1 + \ldots + \xi_n) / n$.

Теорема. Тогда для любого $\varepsilon > 0$

$$\mathbf{P}(\mu - \varepsilon < \bar{\xi} < \mu + \varepsilon) \to 1 \text{ при } n \to \infty.$$

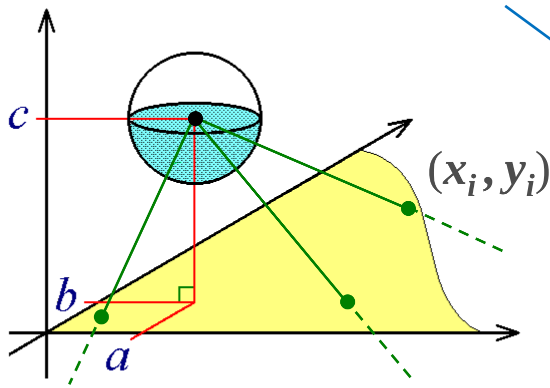
Другими словами, средние арифметические сходятся к математическому ожиданию по вероятности, т. е. при увеличении n распределение $\bar{\xi}$ концентрируется вокруг μ.

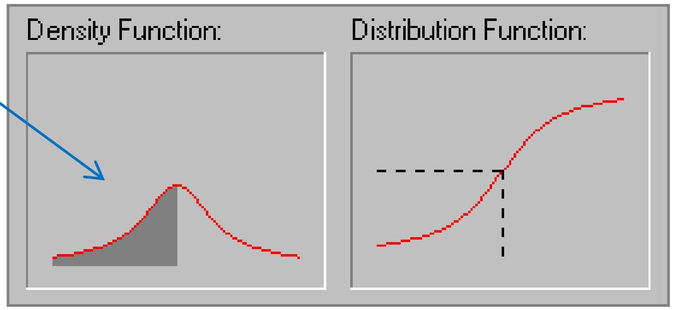


Плотность случайной величины $\bar{\xi}$ стягивается к μ при увеличении n.

Распределение Коши

Контрпример. Пусть $\xi_1, \xi_2, \ldots$ — независимые и одинаково распределенные по закону Коши случайные величины, имеющие плотность $f_\xi(x) = 1/[\pi\,(1 + x^2)]$.





Известно, что, несмотря на симметрию распределения Коши, математическое ожидание сл. в. ξ_1 не существует. Рассмотрим средние арифметические $\overline{\xi} = (\xi_1 + \ldots + \xi_n)\,/\,n$.

Утверждение. $\overline{\xi}$ для любого n распределены так же, как ξ_1. (Следовательно, они не сходятся по вероятности к 0.)

Центральная предельная теорема

Пусть $\xi_1, \xi_2, \ldots$ — независимые и одинаково распределенные случайные величины с математическим ожиданием $\mu = \mathbf{M}\xi_1$ и дисперсией $0 < \sigma^2 = \mathbf{D}\xi_1 < \infty$. Рассмотрим $S_n = \xi_1 + \ldots + \xi_n$. При этом $\mathbf{M}S_n = n\mu$ и $\mathbf{D}S_n = n\sigma^2$.

Теорема. Тогда для любых $a < b$ при $n \to \infty$

$$\mathbf{P}\left(a \le \frac{S_n - \mathbf{M}S_n}{\sqrt{\mathbf{D}S_n}} \le b\right) \to \Phi(b) - \Phi(a),$$

где $\Phi(x)$ — это функция распределения стандартного нормального закона (обозн. $N(0, 1)$), имеющего плотность

$$\varphi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Иначе говоря, распределение центрированных и нормированных сумм S_n сходится к распределению $N(0, 1)$.

График функции распределения закона $N(0,1)$

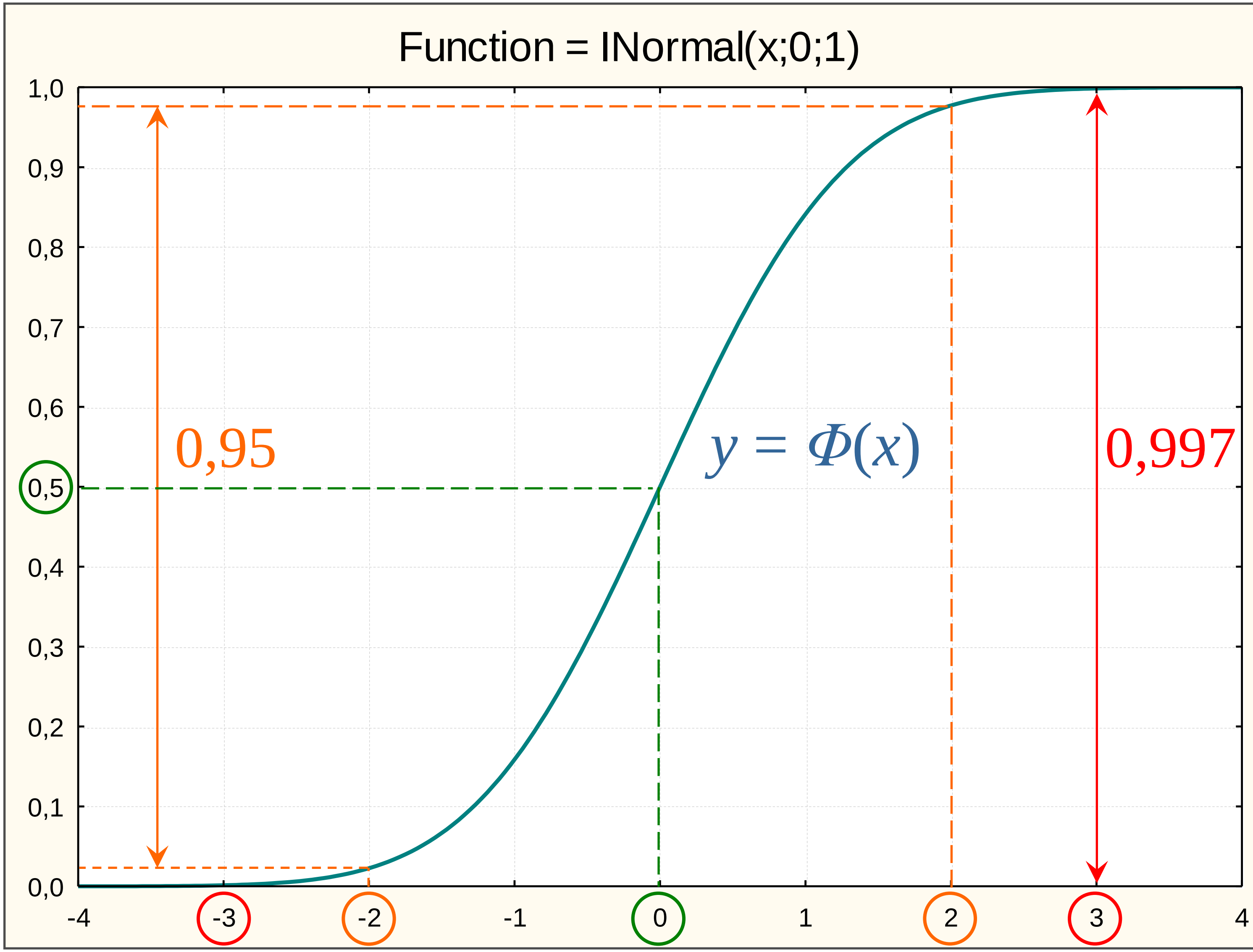
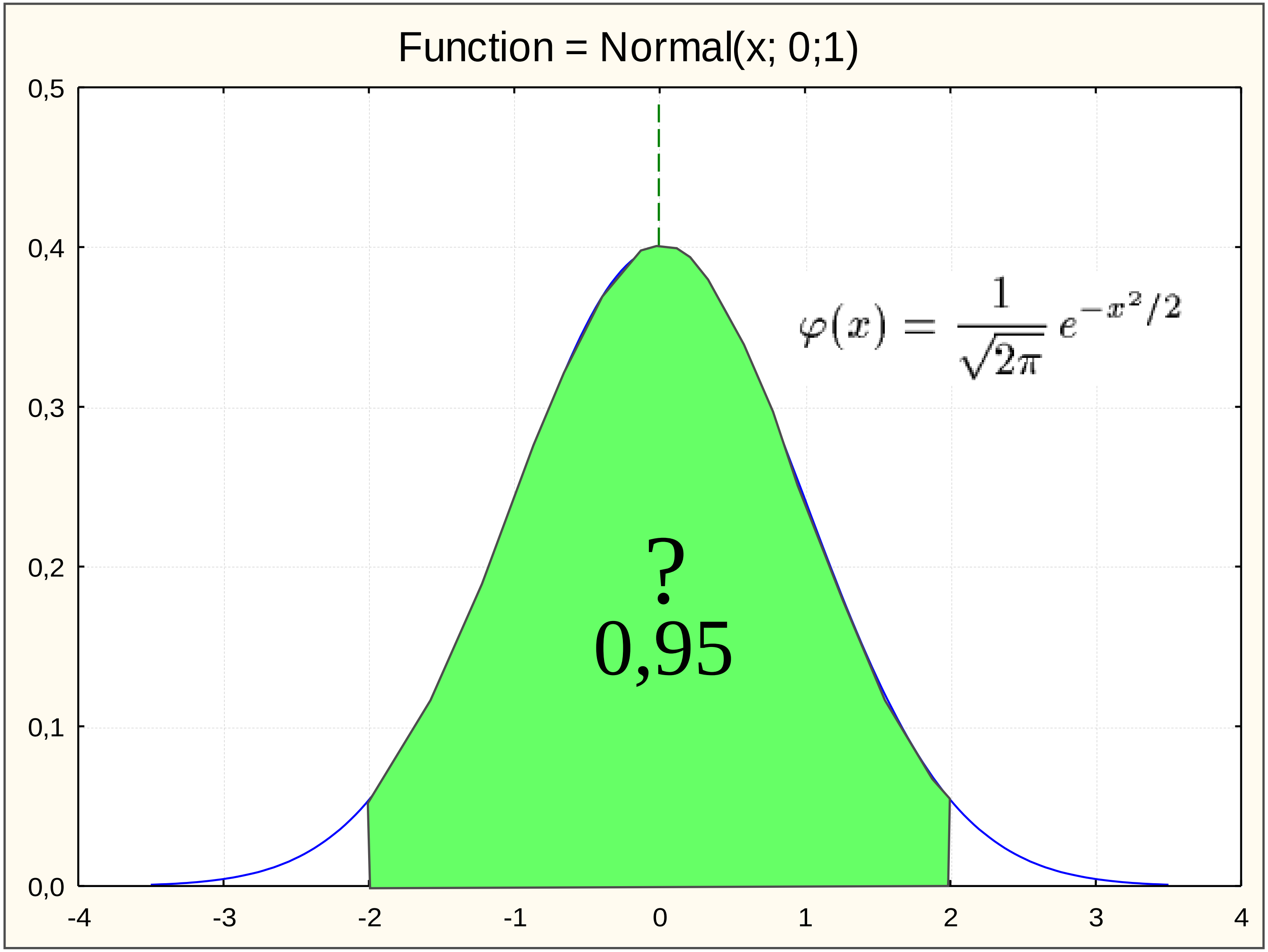


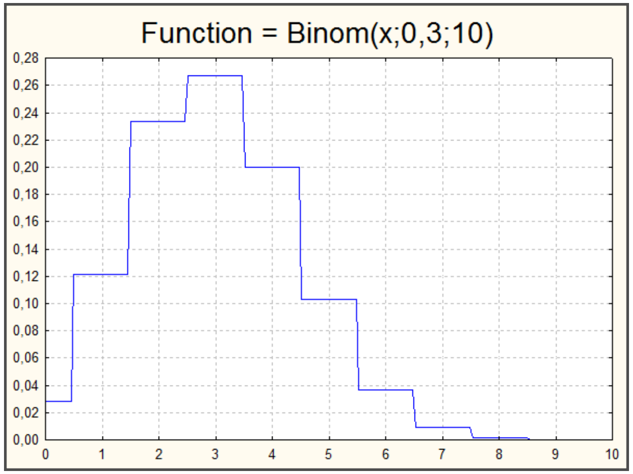
График плотности закона $N(0,1)$

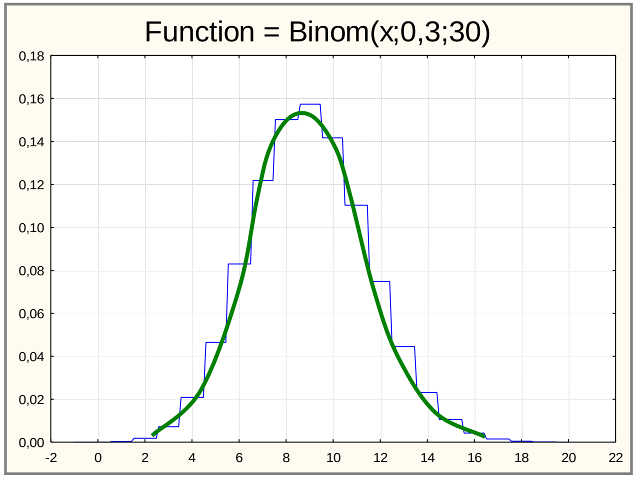


Теорема Муавра — Лапласа

Важный частный случай. Пусть $\xi_1, \xi_2, \ldots$ — независимые случайные величины, имеющие распределение Бернулли:

$$p = \mathbf{P}(\xi_i = 1) = 1 - \mathbf{P}(\xi_i = 0).$$





Какой бы ни была вероятность «успеха» p, при увеличении числа слагаемых n распределение суммы $S_n = \xi_1 + \ldots + \xi_n$ (биномиальное распределение) становится всё более похожим на нормальный закон. На рисунке $p = 0{,}3$; слева приведено распределение для $n = 10$, а справа — для $n = 30$.

Типичная точность оценок

Следствие. Теорема Муавра — Лапласа позволяет найти скорость сходимости частоты $\bar{\xi}$ к вероятности «успеха» p в схеме Бернулли. Нетрудно убедиться, что в данном случае $\mathbf{M}S_n = np$ и $\mathbf{D}S_n = np(1-p)$. Поэтому

$$\frac{S_n - \mathbf{M}S_n}{\sqrt{\mathbf{D}S_n}} = \frac{S_n - np}{\sqrt{np(1-p)}} = \sqrt{n}\,\frac{\bar{\xi} - p}{\sqrt{p(1-p)}} \to N(0,1).$$

Видим, что типичный порядок малости погрешности $|\bar{\xi} - p|$ равен $1/\sqrt{n}$. Действительно, $p(1-p)$ — это константа, а стандартное нормальное распределение $N(0, 1)$ сосредоточено внутри отрезка [–3, 3] с вероятностью 0,997, т. е. его можно считать практически ограниченным.

На следующем слайде приведены графики, которые демонстрируют характер колебаний частоты «успехов» относительно теоретической вероятности $p = 0{,}6$.

Сходимость частоты «успехов» к вероятности в схеме Бернулли

Замечание. Поскольку погрешность частоты имеет порядок малости $1/\sqrt{n}$, для обеспечения трёх верных знаков необходимо иметь выборку размера n = 1 000 000!

Пример применения теоремы Муавра — Лапласа

Вычислим приближённо вероятность, что при $n = 100$ бросаниях правильной монеты число выпавших «гербов» окажется в диапазоне от 35 до 65. Моделью эксперимента служит схема Бернулли с «вероятностью успеха» в отдельном испытании $p = 1/2$.

Пусть S_n — интересующее нас число «успехов». Тогда имеем $\mathbf{M}S_n = np = 50$ и $\mathbf{D}S_n = np(1 - p) = 25$. Отсюда

$$\mathbf{P}(35 \le S_n \le 65) = \mathbf{P}\left(\frac{35-50}{5} \le \frac{S_n - 50}{5} \le \frac{65-50}{5}\right) =$$

$$= \mathbf{P}\left(-3 \le \frac{S_n - 50}{5} \le 3\right) \approx 0{,}997.$$

Практическое задание 2

1) Вычислите приближенно вероятность, что при $n = 100$ бросаниях симметричной монеты число выпавших «гербов» окажется в диапазоне:

а) от 40 до 60,
б) от 30 до 70.

Для этого используйте функцию pnorm, которая вычисляет значения функции распределения $\Phi(x)$ стандартного нормального закона $N(0, 1)$ (название функции pnorm происходит от английских слов probability и normal — вероятность и нормальный)

2) Найдите точно вероятности из а) и б) пункта 1 с помощью функции pbinom, вычисляющей значения функции распределения биномиального закона, т. е. накопленные биномиальные вероятности.

Сравните эти вероятности с результатами из пункта 1.

Теорема Пуассона

Теорема. Пусть в схеме Бернулли $n \to \infty$ и $p \to 0$, причем $np = \lambda > 0$. Тогда $\mathbf{P}(S_n = k) \to p_k = \lambda^k e^{-\lambda}/k!$ при $k = 0, 1, 2, \ldots$. Другими словами, биномиальное распределение сходится к закону Пуассона. Это утверждение иногда называют «законом редких событий».

Доказательство.

$$\mathbf{P}(S_n = k) = \frac{n(n-1)\ldots(n-k+1)}{k!} p^k (1-p)^{n-k} =$$

$$= \frac{(np)^k}{k!}(1-p)^n \left[\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{k-1}{n}\right)(1-p)^{-k}\right],$$

где $(1-p)^n \to e^{-\lambda}$, а выражение в квадратных скобках стремится к 1, поскольку k фиксировано, $n \to \infty$ и $p \to 0$.

Для сравнения, в частности, при $n = 100$ и $p = 0{,}01$ имеем:

k	0	1	2	3	4	5
$\mathbf{P}(S_n = k)$	0,366	0,370	0,185	0,061	0,015	0,003
p_k	0,368	0,368	0,184	0,061	0,015	0,003

О скоростях сходимости в ЦПТ и теореме Пуассона

Следующая теорема содержит оценку для скорости сходимости в центральной предельной теореме.

Теорема Берри — Эссеена. *Пусть* $\mathbf{M}|X_1|^3 < \infty$. *Тогда*

$$\sup_x |F_n(x) - \Phi(x)| \leqslant \frac{C\,\mathbf{M}|X_1 - \mu|^3}{\sigma^3\sqrt{n}} \quad \text{при всех } n,$$

где C *удовлетворяет неравенству* $0{,}399 \approx 1/\sqrt{2\pi} \leqslant C \leqslant 0{,}766$.

Здесь $F_n(x)$ — функция распределения центрированной и нормированной случайной величины S_n. Таким образом, скорость сходимости имеет порядок малости $1/\sqrt{n}$ при $n \to \infty$.

В условиях теоремы Пуассона для любых m и n верна оценка

$$\left| \sum_{k=0}^{m} C_n^k p_n^k (1 - p_n)^{n-k} - e^{-\lambda} \sum_{k=0}^{m} \lambda^k / k! \right| \leq \frac{\lambda^2}{n}.$$

Здесь порядок малости правой части есть $1/n$ при $n \to \infty$.

При формальном построении курса теории вероятностей предельные теоремы появляются в виде своего рода надстройки над элементарными главами теории вероятностей, в которых все задачи имеют конечный, чисто арифметический характер. В действительности, однако, познавательная ценность теории вероятностей раскрывается только предельными теоремами. Более того, без предельных теорем не может быть понято реальное содержание самого исходного понятия всей нашей науки — понятия вероятности.

Б. В. Гнеденко, А. Н. Колмогоров
«Предельные распределения для сумм независимых случайных величин»

Домашнее задание

1) Вычислите приближенно вероятность, что в n = 500 испытаниях Бернулли с вероятностью «успеха» p = 0,0123 число «успехов» будет больше или равно 12.

Для этого используйте функцию ppois, с помощью которой можно вычислять значения функции распределения пуассоновского закона (название функции ppois происходит от английских слов probability и Poisson — вероятность и Пуассон)

2) Найдите точно вероятность из пункта 1 с помощью функции pbinom, вычисляющей значения функции распределения биномиального закона, т.е. накопленные биномиальные вероятности