

Непрерывные случайные величины.
Функция распределения и функция
плотности. Нормальное распределение.
Центральная предельная теорема

Теория вероятностей
и математическая статистика / Урок 4



Непрерывные случайные величины

Ранее мы познакомились с *дискретными* случайными величинами. Такие величины принимают дискретные, т.е. разделимые значения. Разделимость заключается в том, что если случайная величина принимает, например, значения 1 и 2, то она не обязана принимать какие-то промежуточные значения.

Непрерывные случайные величины принимают *все* значения, содержащиеся в заданном промежутке. Промежуток может быть конечным или бесконечным.

Например, рост или вес человека — непрерывные случайные величины: они могут принимать любое значение в некоторых пределах.

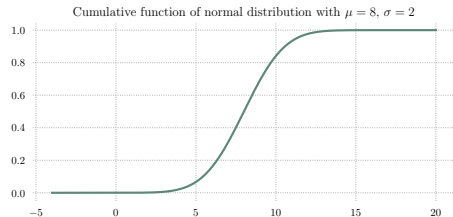
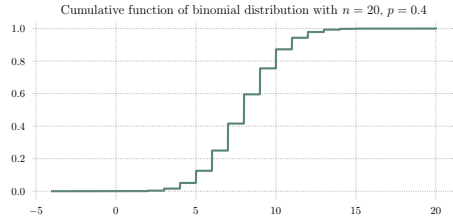
Закон распределения вероятностей дискретной случайной величины мы задавали как соответствие между значениями a_i случайной величины и соответствующими вероятностями $P(X = a_i)$.

Для непрерывных случайных величин аналогичный подход невозможен, поскольку вероятность $P(X = a)$ для непрерывной случайной величины X равна 0 для любого a . Поэтому распределение вероятностей непрерывных случайных величин характеризуют с помощью функции распределения:

$$F(x) = P(X < x)$$

Функция распределения показывает, какова для каждого x вероятность того, что случайная величина X принимает значение меньше x . (Для дискретных распределений эта функция ступенчатая.)

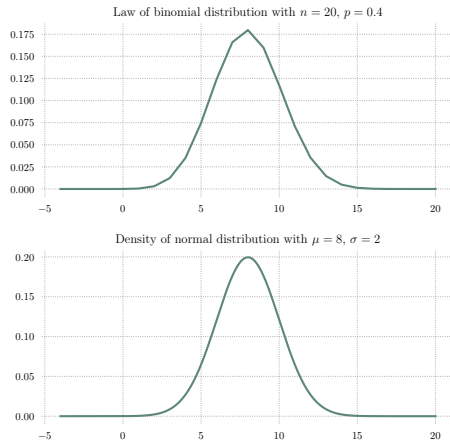
Эта функция монотонно возрастает на отрезке, на котором определена случайная величина. Кроме того, $F(-\infty) = 0$ и $F(\infty) = 1$.



Всё же функция распределения не даёт представления о распределении, аналогичного тому, что даёт закон распределения дискретных случайных величин. Хотелось бы понять, какие значения случайной величины более «вероятно» наблюдать, чем другие.

Для таких целей удобно использовать функцию плотности:

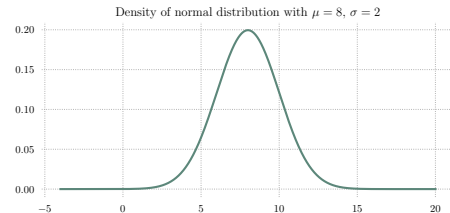
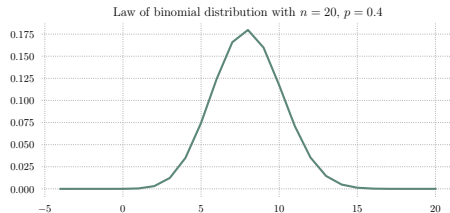
$$f(x) = F'(x)$$



Геометрический смысл функции плотности таков: вероятность того, что случайная величина X будет лежать в отрезке (a, b) , равна площади под графиком функции плотности $f(x)$ в пределах от a до b .

Общая площадь под графиком функции $f(x)$ равна 1, аналогично тому, что сумма вероятностей значений дискретной случайной величины равна 1.

Однако, стоит помнить, что значение $f(x)$ не является вероятностью $P(X = x)$. Оно лишь отражает *плотность* случайной величины в окрестности точки x .



Математическое ожидание и дисперсия для непрерывной случайной величины также считаются иначе, чем для дискретной.

Формула для математического ожидания:

$$M(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Формула для дисперсии:

$$D(X) = \int_{-\infty}^{\infty} (x - M(X))^2 \cdot f(x) dx$$

Примеры непрерывных распределений

Непрерывная случайная величина X имеет нормальное распределение с параметрами μ и $\sigma > 0$, если её плотность распределения задаётся формулой

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

Параметры μ и σ задают, соответственно, математическое ожидание и среднее квадратическое отклонение случайной величины:

$$M(X) = \mu, \quad D(X) = \sigma^2$$

Непрерывная случайная величина X имеет нормальное распределение с параметрами μ и $\sigma > 0$, если её плотность распределения задаётся формулой

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

Параметры μ и σ задают, соответственно, математическое ожидание и среднее квадратическое отклонение случайной величины:

$$M(X) = \mu, \quad D(X) = \sigma^2$$

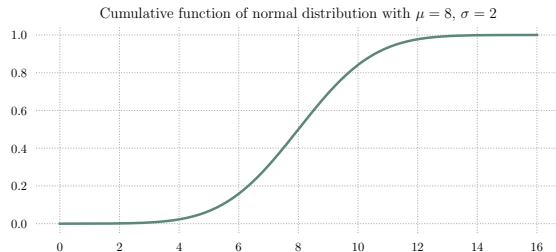
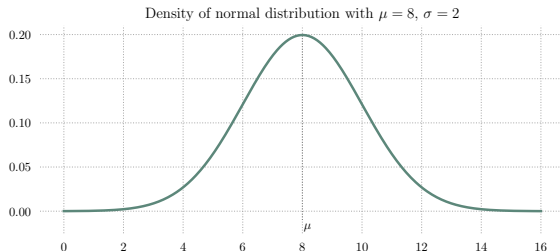
Нормальное распределение с параметрами $\mu = 0$ и $\sigma = 1$ называется стандартным нормальным распределением.

Нормальное распределение является одним из наиболее распространённых на практике. Например, нормально распределены:

- рост, вес людей,
- показатели IQ,
- время прихода на работу,
- скорость движения молекул в жидкостях и газах.

Как правило, нормально распределёнными являются случайные величины, описывающие события, которые зависят от большого числа слабо связанных случайных факторов.

Так выглядят графики плотности нормального распределения и функции нормального распределения.



В модуле `stats` из библиотеки `scipy` содержатся реализации основных функций для различных распределений:

- ① `binom` — биномиальное,
- ② `poisson` — Пуассоновское,
- ③ `geom` — геометрическое,
- ④ `norm` — нормальное,
- ⑤ `uniform` — непрерывное равномерное

и др.

Доступные функции:

- ① pmf — закон распределения для дискретных величин,
- ② pdf — функция плотности для непрерывных величин,
- ③ cdf — функция распределения,
- ④ ppf — квантильная функция (обратная к функции распределения)

и др.

Доступные функции:

- ① pmf — закон распределения для дискретных величин,
- ② pdf — функция плотности для непрерывных величин,
- ③ cdf — функция распределения,
- ④ prf — квантильная функция (обратная к функции распределения)

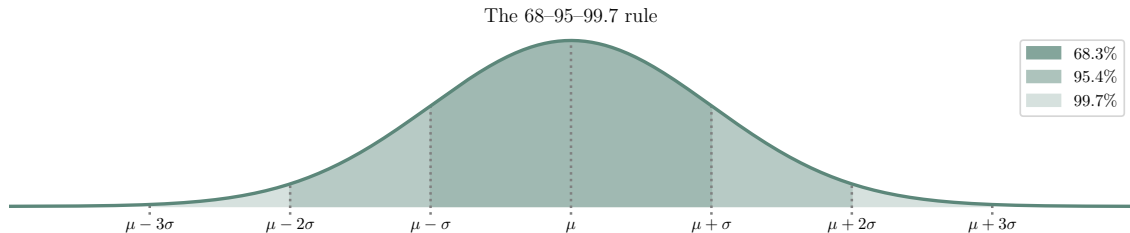
и др.

Например,

- Функция плотности нормального распределения с параметрами $\mu = 8$, $\sigma = 2$:
`stats.norm.pdf(x, loc=8, scale=2)`
- Можно сразу же зафиксировать распределение:
`norm = stats.norm(loc=8, scale=2),`
а затем пользоваться различными функциями напрямую:
`norm.pdf(x), norm.cdf(x)` и др.

Для вычисления разброса значений нормально распределённой случайной величины можно использовать следующие правила:

- Интервал от $\mu - \sigma$ до $\mu + \sigma$ (стандартное отклонение) содержит около 68% вероятностной массы (т.е. с вероятностью 68% данная величина попадает в этот интервал).
- От $\mu - 2\sigma$ до $\mu + 2\sigma$ — около 95% массы (правило двух сигм).
- От $\mu - 3\sigma$ до $\mu + 3\sigma$ — около 99.7% массы (правило трёх сигм).



- Экспоненциальное (или показательное) распределение (`scipy.stats.expon`): время между последовательными свершениями одного и того же события. Является непрерывным аналогом геометрического распределения. Функция плотности:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

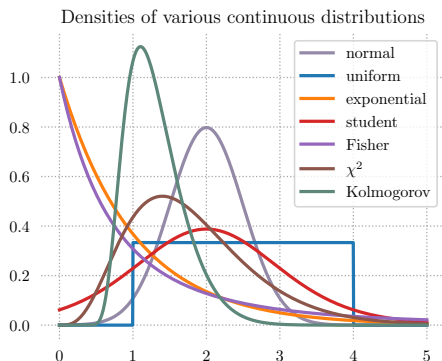
- Непрерывное равномерное распределение (`scipy.stats.uniform`) — непрерывный аналог дискретного равномерного распределения. Функция плотности постоянна внутри отрезка:

$$f(x) = \begin{cases} 1/(b-a), & x \in [a, b], \\ 0 & \text{иначе.} \end{cases}$$

Упомянем также несколько распределений, на которых построены некоторые статистические методы, про которые мы будем говорить в будущем:

- Распределение Стьюдента (`scipy.stats.t`)
- Распределение Фишера (`scipy.stats.f`)
- Распределение χ^2 (хи-квадрат, `scipy.stats.chi2`)
- Распределение Колмогорова (`scipy.stats.ksone`)

Кстати, про распределение Стьюдента (оно нам ещё не раз пригодится в будущем) есть раздел в дополнительных материалах к уроку.

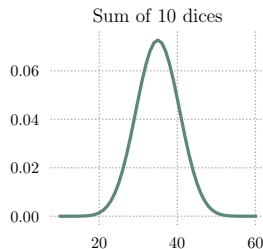
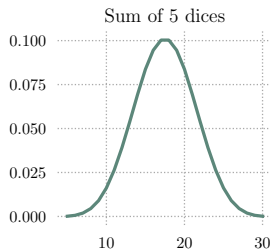
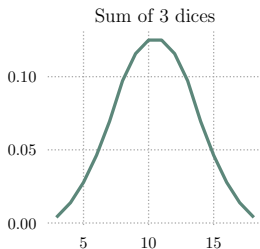
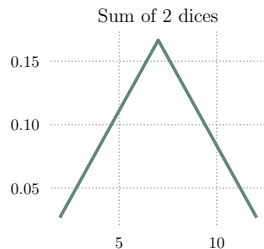


Центральная предельная теорема

Нормальное распределение обладает свойством **устойчивости**. Это означает, что с нормальными распределениями можно проводить различные арифметические операции (такие как сложение, вычитание, умножение на константы), и нормальное распределение останется нормальным.

Большинство других распределений таковыми не являются. Например, на втором уроке мы рассматривали распределение суммы значений двух подбрасываемых игральных кубиков. Распределение значений каждого кубика отдельно является дискретным равномерным, но их сумма уже не имеет равномерное распределение.

Оказывается, неустойчивые распределения к чему-то стремятся. Например, ниже изображены законы распределения сумм разного числа кубиков. На что это становится похоже?



Да, если суммировать неустойчивые распределения, результат становится всё более похож на нормальное распределение. В этом и заключается **Центральная предельная теорема**.

Пусть имеется n случайных величин X_1, \dots, X_n , имеющих одинаковое распределение с математическим ожиданием M и дисперсией D . Пусть $Y = X_1 + \dots + X_n$ — сумма этих случайных величин.

Центральная предельная теорема утверждает: чем больше n , тем ближе распределение величины Y к нормальному распределению с параметрами

$$\mu = n \cdot M, \quad \sigma^2 = n \cdot D$$

Пусть имеется n случайных величин X_1, \dots, X_n , имеющих одинаковое распределение с математическим ожиданием M и дисперсией D . Пусть $Y = X_1 + \dots + X_n$ — сумма этих случайных величин.

Центральная предельная теорема утверждает: чем больше n , тем ближе распределение величины Y к нормальному распределению с параметрами

$$\mu = n \cdot M, \quad \sigma^2 = n \cdot D$$

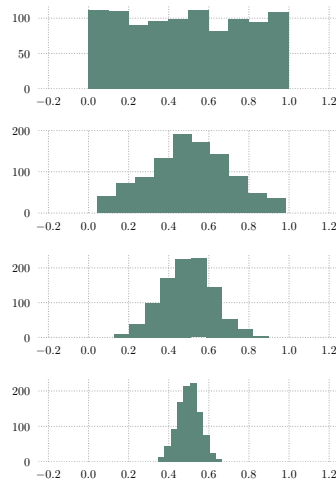
Другая версия этой теоремы: пусть $Z = \frac{1}{n} \sum_{i=1}^n X_i$ — среднее арифметическое случайных величин. Тогда с увеличением n распределение этой величины становится всё ближе к нормальному распределению с параметрами

$$\mu = M, \quad \sigma^2 = D/n$$

Продemonстрируем центральную предельную теорему на примере равномерного распределения. В каждом случае 1000 раз берётся выборка из n значений случайной величины, считается среднее значений из выборки, затем строится гистограмма распределения этого среднего.

Число n равно, соответственно, 1 (в этом случае имеем обычное равномерное распределение), 2, 5 и 30.


При $n = 30$ распределение уже не отличить от нормального.



В реальных ситуациях у нас нет возможности генерировать большое количество выборок. Как правило, у нас есть только одна. Вычисляя, например, выборочное среднее, мы получим только одно значение. Зачем тогда нужна центральная предельная теорема?

ЦПТ позволяет пролить свет на распределение этого самого выборочного среднего. Т.е. благодаря ЦПТ, мы знаем, какие значения выборочного среднего можно ожидать, будь у нас ещё одна или несколько выборок.

Сравнение ожидаемых и наблюдаемых значений является ключевым моментом при проверке статистических гипотез, о которых мы поговорим на следующем уроке.



Спасибо за внимание