Теория вероятностей и математическая статистика

Вебинар 4

Непрерывные случайные величины. Функция распределения
и плотность распределения. Равномерное и нормальное распределение.
Центральная предельная теорема

Непрерывные случайные величины

Непрерывные случайные величины

Ранее мы познакомились с *дискретными* случайными величинами. Такие величины принимают дискретные, т.е. разделимые значения. Например, это может быть конечное или счётное множество значений.

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.

Функция распределения

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

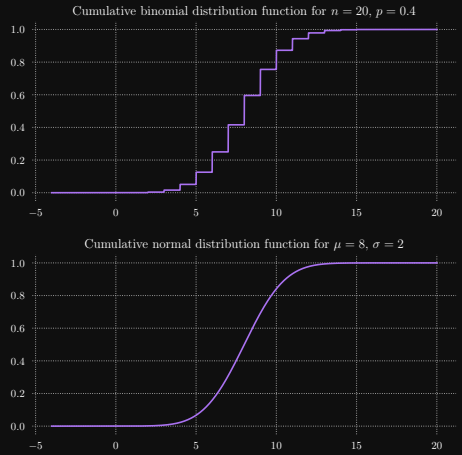
Для непрерывных случайных величин аналогичный подход невозможен, поскольку вероятность $P(X = a)$ для непрерывной случайной величины X равна 0 для любого a. Поэтому распределение вероятностей непрерывных случайных величин характеризуют с помощью *функции распределения*:

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

Функция распределения

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 $F(\infty) = 1$.

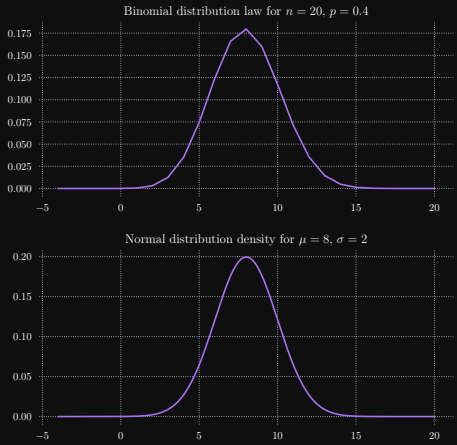


Плотность распределения

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*:

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

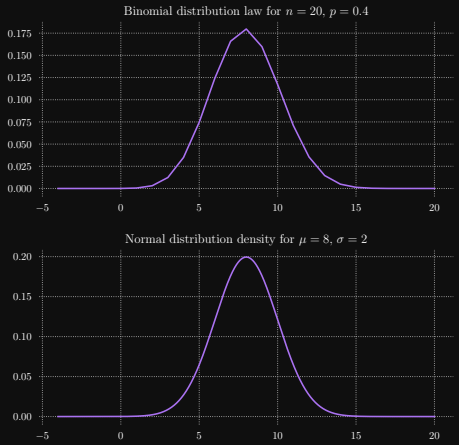


Плотность распределения

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

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

Равномерное распределение

Непрерывная случайная величина X имеет *равномерное распределение* на отрезке $[a, b]$, если её плотность внутри этого отрезка постоянна, а вне этого отрезка равна 0. Другими словами:

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

Не путать с *дискретным* равномерным распределением.

Равномерное распределение

Непрерывная случайная величина X имеет *равномерное распределение* на отрезке $[a, b]$, если её плотность внутри этого отрезка постоянна, а вне этого отрезка равна 0. Другими словами:

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

Не путать с *дискретным* равномерным распределением.

Математическое ожидание и дисперсия равномерного распределения:

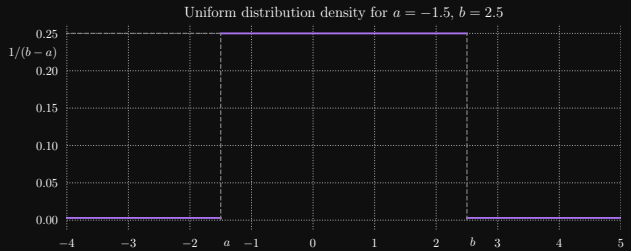
$$M(X) = \frac{a+b}{2}, \ D(X) = \frac{(b-a)^2}{12}$$

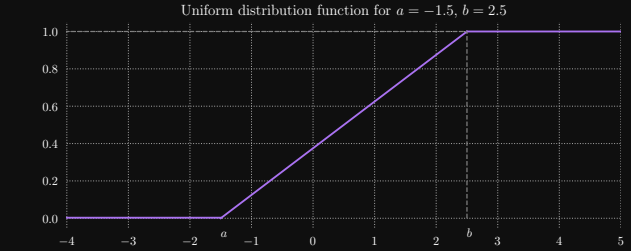
Равномерное распределение

Так выглядят графики плотности равномерного распределения и функции равномерного распределения.

Кстати, формула функции равномерного распределения:

$$F(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & x \in [a, b], \\ 1, & x > b. \end{cases}$$





Нормальное распределение

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:

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

Нормальное распределение

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:

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

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

Нормальное распределение

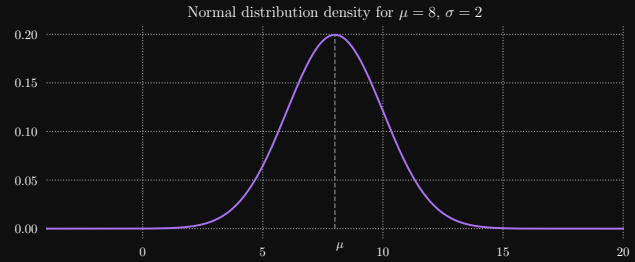
Нормальное распределение является одним из наиболее распространённых на практике. Например, нормально распределены:

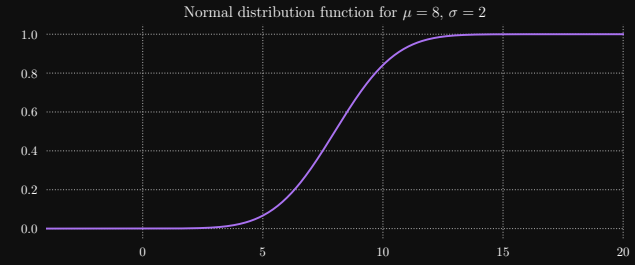
- рост, вес людей,
- показатели IQ,
- время прихода на работу,
- скорость движения молекул в жидкостях и газах.

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

Нормальное распределение

Так выглядят графики плотности нормального распределения и функции нормального распределения.





Нормальное распределение

Функция нормального распределения:

$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right],$$

где erf — *функция ошибок*.

Функция ошибок представляет собой интеграл

$$\operatorname{erf}(x) = \frac{2}{\pi}\int_0^x e^{-t^2}dt,$$

который аналитически не считается.

Нормальное распределение

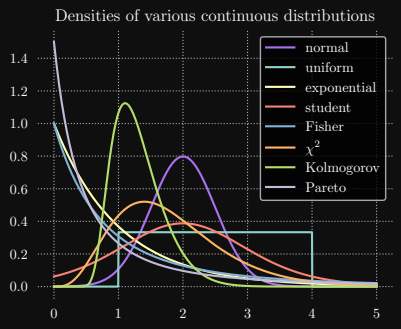
Для вычисления разброса значений нормально распределённой случайной величины можно использовать следующие правила:

- 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*).

Другие непрерывные распределения

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

$$f(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$



- t-распределение Стьюдента.
- Распределение Фишера.
- Распределение χ^2 (хи-квадрат).
- Распределение Колмогорова.
- Распределение Парето («Правило 20 к 80»).

Центральная предельная теорема

Устойчивость

Одно из практически уникальных свойств нормального распределения — *устойчивость* — означает, что если X и Y — *независимые нормально распределённые* случайные величины, то их комбинация $Z = a \cdot X + b \cdot Y$ (a, b — числа) также имеет нормальное распределение. Более того, для распределения Z верны следующие равенства. Математическое ожидание:

$$M(Z) = a \cdot M(X) + b \cdot M(Y)$$

Дисперсия:

$$D(Z) = |a| \cdot D(X) + |b| \cdot D(Y)$$

Устойчивость

Одно из практически уникальных свойств нормального распределения — *устойчивость* — означает, что если X и Y — *независимые нормально распределённые* случайные величины, то их комбинация $Z = a \cdot X + b \cdot Y$ (a, b — числа) также имеет нормальное распределение. Более того, для распределения Z верны следующие равенства. Математическое ожидание:

$$M(Z) = a \cdot M(X) + b \cdot M(Y)$$

Дисперсия:

$$D(Z) = |a| \cdot D(X) + |b| \cdot D(Y)$$

Большинство других распределений не являются устойчивыми. Например, сумма двух равномерно распределённых случайных величин не является равномерно распределённой. Вместо этого неустойчивые распределения «стремятся» к нормальному. Это хорошо иллюстрирует центральная предельная теорема.

Центральная предельная теорема

Рассмотрим выборку из n значений случайной величины X, имеющей произвольное распределение, и пусть Y — случайная величина, равная сумме этих значений.

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

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

Центральная предельная теорема

Рассмотрим выборку из n значений случайной величины X, имеющей произвольное распределение, и пусть Y — случайная величина, равная сумме этих значений.

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

$$\mu = n \cdot M(X), \;\; \sigma^2 = n \cdot D(X)$$

Другая версия этой теоремы: пусть Z — случайная величина, равная среднему арифметическому значений из выборки. Тогда с увеличением n распределение этой величины становится всё ближе к нормальному распределению с параметрами

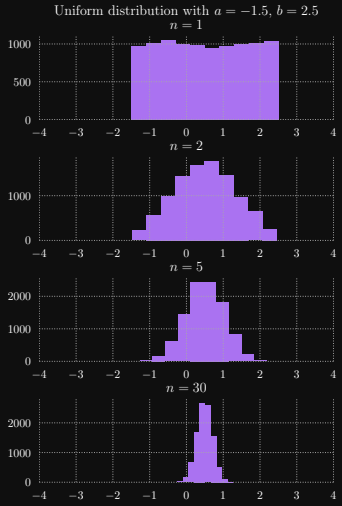
$$\mu = M(X), \;\; \sigma^2 = \frac{D(X)}{n}$$

Центральная предельная теорема

Центральная предельная теорема согласуется со сделанным ранее наблюдением, что, как правило, случайные величины, описывающие события, которые зависят от большого числа слабо связанных случайных факторов, являются нормально распределёнными.

Продемонстрируем центральную предельную теорему на примере равномерного распределения с параметрами $a = -1.5$ и $b = 2.5$. В каждом случае 10000 раз берётся выборка из n значений случайной величины, считается среднее значений из выборки, затем строится гистограмма распределения этого среднего.

Отметим, что центральная предельная теорема работает не только для непрерывных случайных величин, но и для дискретных.



На следующем занятии

Проверка статистических гипотез.
P-значения. Доверительные интервалы