Глоссарий

Распределение случайной величины — описание того, с какой вероятностью случайная величина может принять те или иные значения.

Для дискретной случайной величины это вероятности всех значений, которые величина может принимать.

Для непрерывной — функция плотности вероятности, с помощью которой можно найти вероятность того, что случайная величина попадёт в произвольный промежуток. Обычно функция плотности вероятности обозначается f(x).

Носитель распределения — множество значений случайной величины, для которых вероятность не равна нулю.

Эксперимент Бернулли — эксперимент с двумя исходами, которые часто обозначают как успех и неудача. (см. ниже «Распределение Бернулли» в разделе «Распределения»)

Биномиальный эксперимент, или схема Бернулли — независимые повторения (часто говорят — испытания) одного и того же биномиального эксперимента конечное число раз. (см. ниже «Биномиальное распределение» в разделе «Распределения»)

CDF, или функция распределения, или кумулятивная функция распределения — функция, которая для заданного распределения случайной величины показывает вероятность того, что эта случайная величина примет значение не больше, чем аргумент этой функции: $F(x) = P(X \leq x)$

Для дискретных случайных величин это сумма вероятностей всех значений, меньше или равных x. В этом случае можно записать так: $F(x) = P(X \leq x) = = \sum_{x_i \leq x} P(X = x_i)$

Для непрерывных — вероятность попадания в промежуток от ∞ до x или, что то же самое, от минимального значения, для которого вероятность не равна нулю, до x. То есть $F(x) = P(X \leq x)$ — это площадь под графиком функции плотности вероятности на промежутке от ∞ до x.

PPF — функция, обратная CDF: по заданной вероятности того, что случайная величина окажется меньше некоторого значения, находит это значение.

Аппроксимация — нахождение для некоторого распределения A другого распределения B, чаще всего другого типа, по которому можно найти приблизительные вероятности того, что случайная величина, имеющая распределение A, примет те или иные значения. (см. ниже раздел «Аппроксимации»)

Распределения

Распределение Бернулли

Используется для **дискретных** случайных величин.

Носитель (для каких значений вероятность не равна нулю): 0 и 1.

Описание: Распределение вероятностей успеха и неудачи в эксперименте Бернулли.

Параметр: p — вероятность успеха.

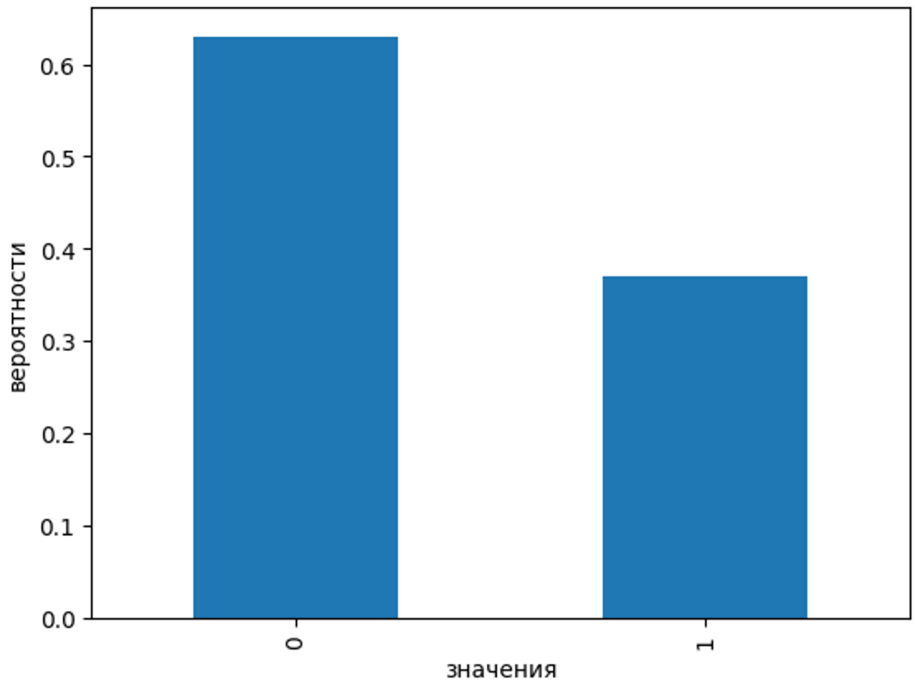
Математическое ожидание: p

Дисперсия: $p \cdot (1 - p)$

Пример визуализации в python:

```
import pandas as pd

p = 0.37 # вероятность успеха

probabilities = pd.Series([1 - p, p])

# отобразим график распределения
probabilities.plot.bar(xlabel='значения', ylabel='вероятности')
```

Результат выполнения кода:



Биномиальное распределение

Используется для **дискретных** случайных величин.

Носитель: целые числа от 0 до n.

Описание: распределение вероятностей количества успехов в биномиальном эксперименте.

Параметры: n — количество независимых повторений эксперимента Бернулли, p — вероятность успеха в каждом из них.

Вероятность того, что случится ровно k успехов из n испытаний: $P(X = k) = C_n^k \cdot p^k \cdot (1-p)^{n-k}$, где C_n^k — число сочетаний.

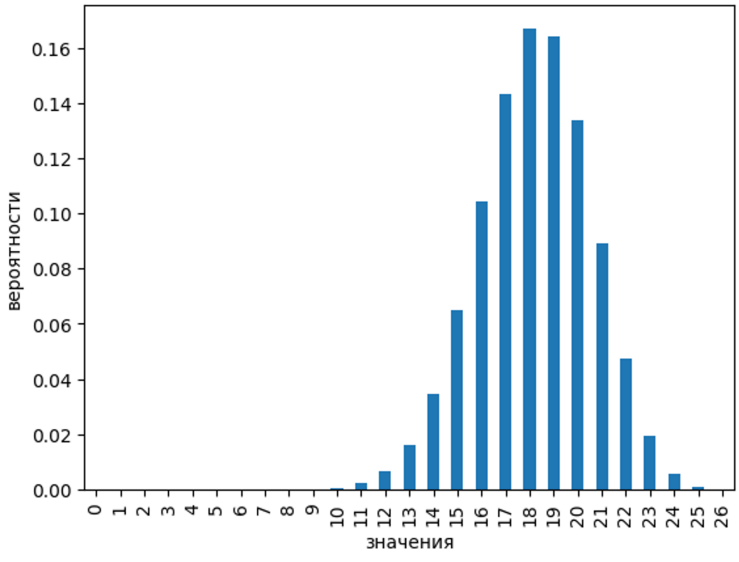
Математическое ожидание: $n \cdot p$

Дисперсия: $n \cdot p \cdot (1-p)$

Пример визуализации в python:

```
import pandas as pd
from scipy.stats import binom

n = 26 # количество попыток
p = 0.7 # вероятность успеха

distr = pd.Series()

# найдём вероятности для всех возможных значений — от нуля до n
for k in range(0, n + 1):
    distr[k] = binom.pmf(k, n, p)

# отобразим график распределения
distr.plot.bar(xlabel='значения', ylabel='вероятности')
```

Результат выполнения кода:



Распределения

Равномерное непрерывное распределение

Используется для **непрерывных** случайных величин.

Носитель: конечный промежуток между двумя числами.

Описание: распределение вероятности того, что непрерывная случайная величина примет с одинаковой вероятностью любое значение между двумя числами.

Параметры: a — левая граница промежутка, b — правая граница.

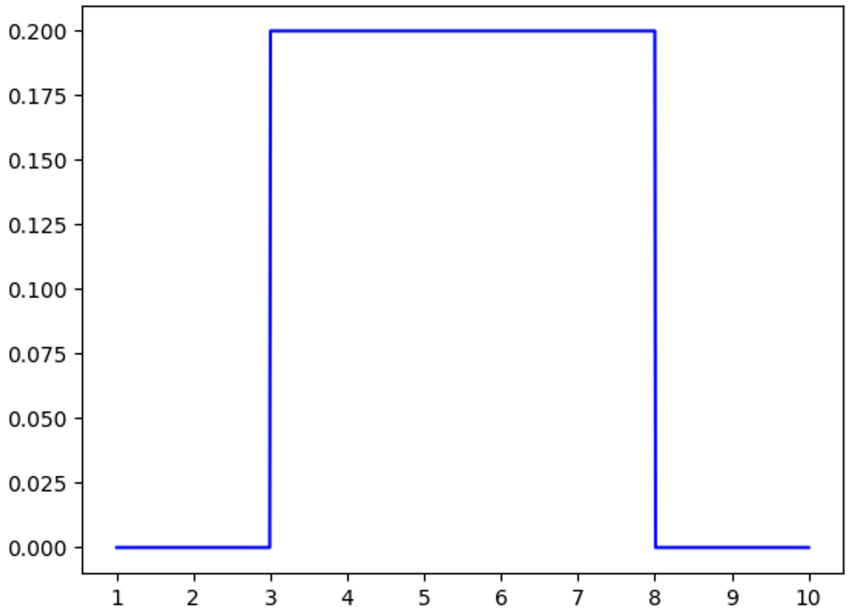
Математическое ожидание: середина промежутка между a и b, то есть $(a+b)/2$

Дисперсия: $(b-a)^2/12$

Пример визуализации в python:

```
import scipy.stats as st
import matplotlib.pyplot as plt

# параметры распределения
a = 3
b = 8

# зададим функцию, которая выдаёт <num> чисел,
# равномерно распределённых от <start> до <stop>
# (понадобится для визуализации распределения)
def linspace(start, stop, num):
    step = (stop - start) / (num - 1)
    result = []
    for i in range(num):
        result.append(start + step * i)
    return result

# зададим отступ влево и вправо от параметров a и b для визуализации
margin = 2

# получим список из 1000 чисел от a-отступ до b+отступ:
# в этих границах будет отображаться график
x = linspace(a - margin, b + margin, 1000)

# зададим все целые числа в пределах промежутка a-отступ до b+отступ
# как подписи на горизонтальной оси
theplot = plt.subplot()
x_ticks = linspace(a - margin, b + margin, b - a + 2 * margin + 1)
theplot.set_xticks(x_ticks)

# отобразим график распределения
plt.plot(x, st.uniform.pdf(x, loc=a, scale=b-a), color="blue");
```

Результат выполнения кода:



Нормальное распределение

Используется для **непрерывных** случайных величин.

Носитель: вся числовая ось.

Описание: Симметричное распределение вероятностей случайной величины, наиболее вероятно принимающей значения ближе к центру распределения и с меньшей вероятностью далеко от него.

Параметры: μ — центр распределения, σ — стандартное отклонение распределения.

Математическое ожидание: μ

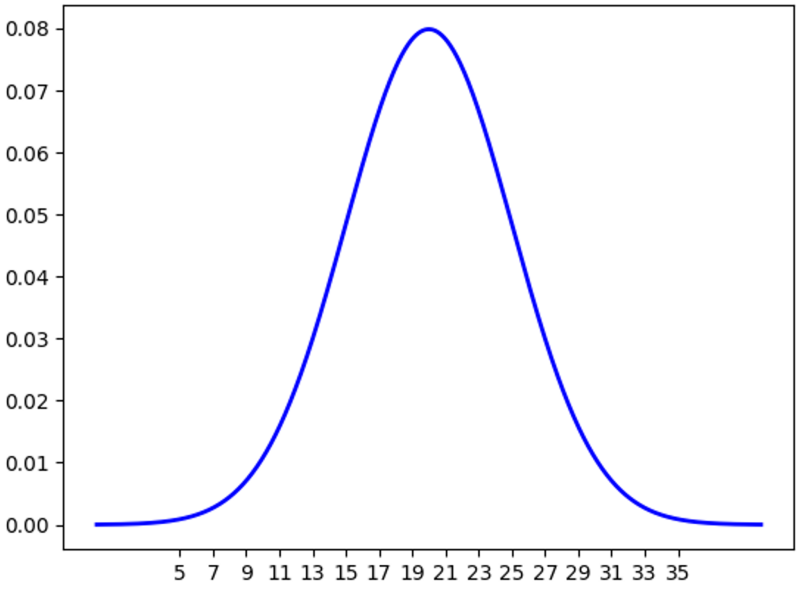
Дисперсия: σ^2

Распределения

Пример визуализации в python:

```
import matplotlib.pyplot as plt
from scipy.stats import norm

# зададим функцию, которая выдаёт <num> чисел,
# равномерно распределённых от <start> до <stop>
# (понадобится для визуализации распределения)
def linspace(start, stop, num):
    step = (stop - start) / (num - 1)
    result = []
    for i in range(num):
        result.append(start + step * i)
    return result

# задаём параметры нормального распределения
mu = 20
sigma = 5

# границы промежутка, на котором будем строить график: mu плюс-минус 4 sigma
left = int(mu - 4 * sigma)
right = int(mu + 4 * sigma)

# рисуем график нормального распределения:
# получим список из 1000 чисел от left до right
x = linspace(left, right, 1000)
# используем синюю линию ширины 2
theplot = plt.subplot()
theplot.plot(x, norm.pdf(x, mu, sigma), 'b-', lw=2)

# зададим подписи по оси x в пределах ± 3 ст.отклонений от мат.ожидания
x_ticks = linspace(int(mu - 3 * sigma), int(mu + 3 * sigma), 3 * sigma + 1)
theplot.set_xticks(x_ticks)

# отобразим график распределения
plt.show()
```

Результат выполнения кода:



Распределение Пуассона

Используется для **дискретных** случайных величин.

Носитель: неотрицательные целые числа.

Описание: распределение вероятностей количества событий за фиксированный промежуток времени для процесса, который происходит с заданной интенсивностью, то есть известно, сколько обычно происходит событий за этот промежуток времени.

Параметр: λ — интенсивность процесса.

Вероятность того, что случится ровно k событий в течение фиксированного промежутка времени: $P(X = k) = (\lambda^k \cdot e^{-\lambda})/k!$

Математическое ожидание: λ

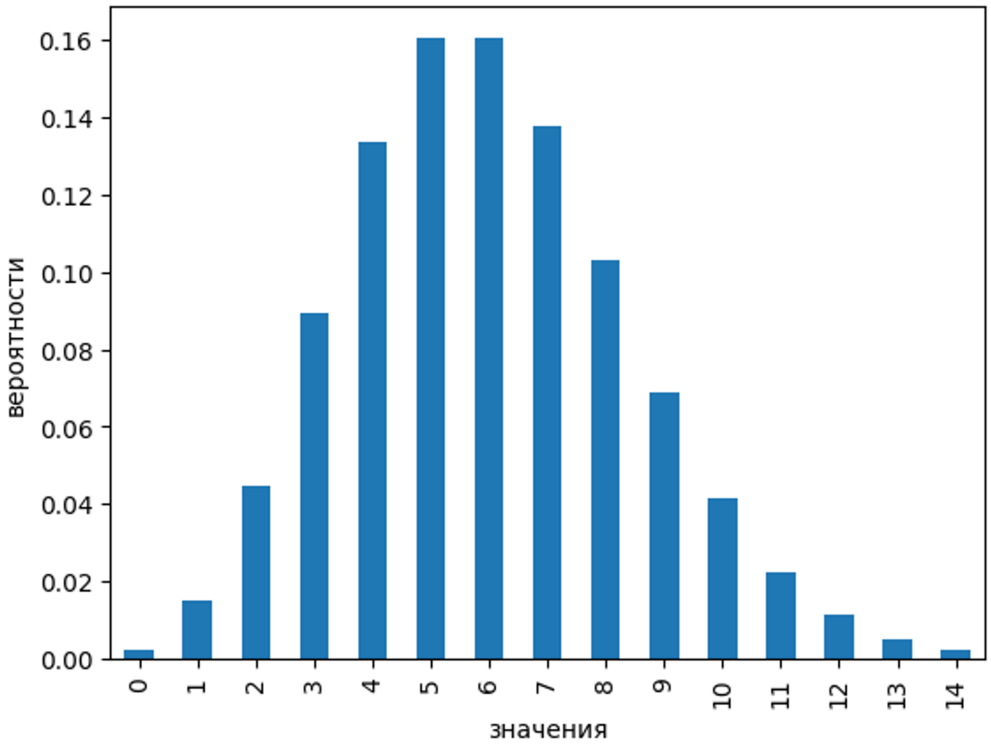
Дисперсия: λ

Пример визуализации в python:

```
import pandas as pd
import numpy as np
from scipy.stats import poisson

lmbd = 6 # параметр лямбда

distr = pd.Series()

# найдём значения в промежутке плюс-минус 4 ст.отклонения от мат.ожидания,
# при этом левая граница не может быть меньше нуля:
left = np.maximum(0, lmbd - 4 * sqrt(lmbd))
right = lmbd + 4 * sqrt(lmbd)
for k in range(int(left), int(right)):
    distr[k] = poisson.pmf(k, lmbd)

distr.plot.bar(xlabel='значения', ylabel='вероятности')
```

Распределения

Результат выполнения кода:



Распределение Пуассона

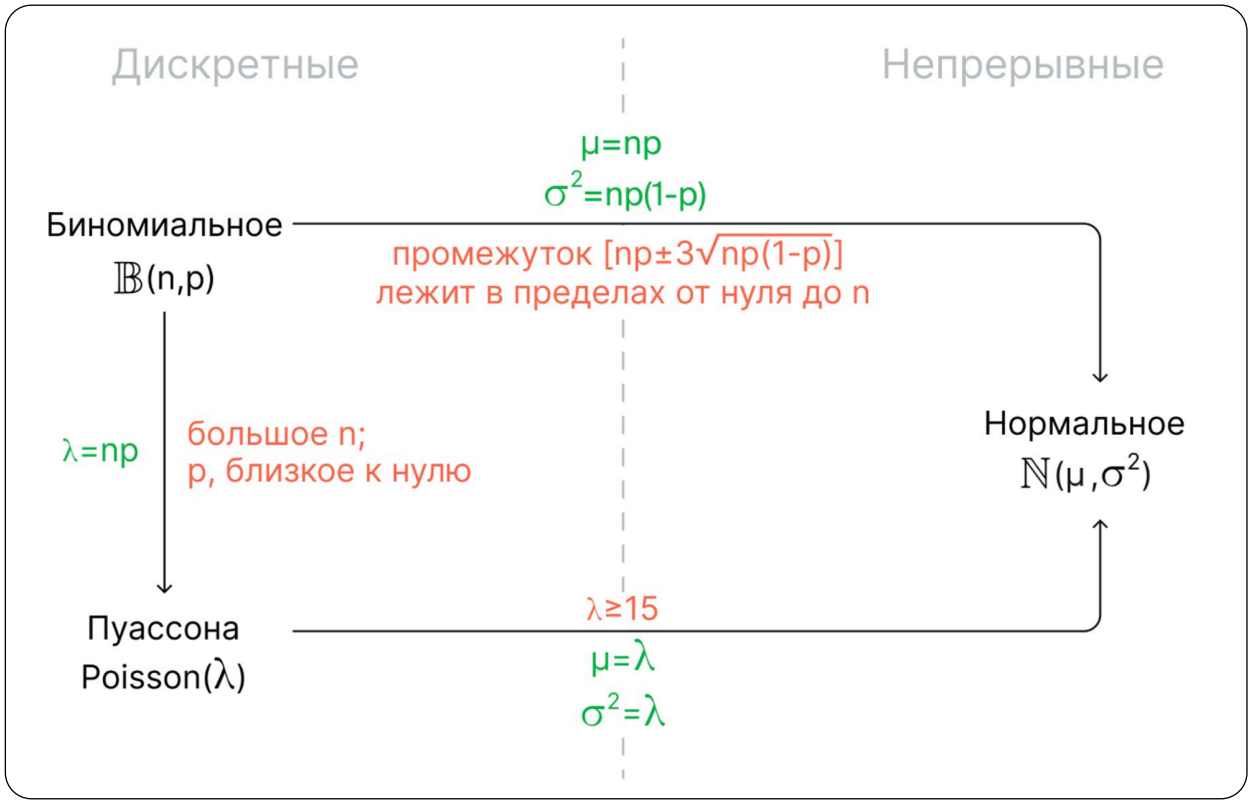
Аппроксимации

Общая схема рассмотренных аппроксимаций

Направление стрелок: что аппроксимируем → чем аппроксимируем.

Красным указаны условия применимости аппроксимации.

Зелёным — параметры аппроксимирующего распределения, они зависят от параметров аппроксимируемого распределения.



Аппроксимация биномиального распределения распределением Пуассона

Условие аппроксимации

Пусть X— случайная величина, имеющая биномиальное распределение: $X \sim \mathbb{B}(n, p)$. Тогда если p достаточно мало, а n достаточно велико, X может быть аппроксимирована распределением Пуассона с параметром $\lambda = n \cdot p$.

Применение аппроксимации

Случайная величина X дискретна, её носитель — целые числа от 0 до n. Аппроксимирующая случайная величина тоже дискретна, её носитель — все целые числа.

Значит, для каждого возможного значения X можно рассчитать вероятность того, что это значение примет аппроксимирующая случайная величина. Это и будет приблизительная вероятность для любого возможного значения X.

Ошибка аппроксимации

Для уменьшения ошибки аппроксимации важно, чтобы параметр p был близок к нулю, а параметр n был большим. Условно, для хорошей аппроксимации достаточно, чтобы n измерялось в сотнях или больше, а p — в единицах процентов или меньше.
Увеличение n в два раза даёт практически такое же уменьшение ошибки, как и уменьшение p в два раза. Это касается и средней, и максимальной ошибок, — то есть и средней разницы между аппроксимируемыми и аппроксимирующими значениями, и максимальной такой разницы для всех значений от 0 до n.
Подробнее об этом можно посмотреть под катом в конце десятого урока в этой теме.

Аппроксимация биномиального распределения нормальным

Условие аппроксимации

Пусть X— случайная величина, имеющая биномиальное распределение: $X \sim \mathbb{B}(n, p)$.

Если промежуток от математического ожидания биномиального распределения $n \cdot p$ плюс-минус три его стандартных отклонения $\sqrt{n \cdot p \cdot (1-p)}$ лежит в пределах $[0, n]$то аппроксимацию можно использовать.

В таком случае X можно аппроксимировать нормальным распределением с параметрами $\mu = n \cdot p$ и $\sigma = \sqrt{n \cdot p \cdot (1-p)}$

👉 Часто условиями аппроксимации называют большое n и близость p к 0.5. Как мы пояснили в уроке об аппроксимации биномиального распределения нормальным, такое условие неполное: оно охватывает не все случаи, когда аппроксимация работает.

Применение аппроксимации

X— дискретная случайная величина, а аппроксимирующая её случайная величина имеет нормальное распределение, то есть непрерывна.

Вероятность того, что непрерывная случайная величина примет любое конкретное значение, равна нулю. Для неё можно найти только вероятность попадания в промежуток. Поэтому необходимо использовать поправку на непрерывность: для каждого значения искать вероятность того, что аппроксимирующая случайная величина попадёт в промежуток $[x_i - 0.5, x_i + 0.5]$.

Ошибка аппроксимации

Чем симметричнее биномиальное распределение, тем точнее его аппроксимирует нормальное.
Верно следующее:

- Чем ближе p к 0.5, тем меньше ошибки аппроксимации, — и средняя, и максимальная. То есть и средняя разница между аппроксимируемыми и аппроксимирующими значениями, и максимальная такая разница для всех значений от 0 до n.
- Чем ближе p к 0.5, тем меньше должно быть n, чтобы условие аппроксимации выполнялось.
- Чем больше n, тем меньше ошибки аппроксимации, — и средняя, и максимальная — для одного и того же p.

Подробнее об этом можно посмотреть под катом в конце одиннадцатого урока в этой теме.

Аппроксимация распределения Пуассона нормальным

Условие аппроксимации

Пусть X— случайная величина, имеющая распределение Пуассона:
$X \sim Poisson(\lambda)$.

Если $\lambda \geq 15$, X может быть аппроксимирована нормальным распределением с параметрами $\mu = \lambda, \sigma = \sqrt{\lambda}$.

🤝 Эта граница конвенциональна: иногда указывают $\lambda \geq 10$.

Применение аппроксимации

X— дискретная случайная величина, а аппроксимирующая её случайная величина имеет нормальное распределение, то есть непрерывна. Вероятность того, что непрерывная случайная величина примет любое конкретное значение, равна нулю. Для неё можно найти только вероятность попадания в промежуток. Поэтому необходимо использовать поправку на непрерывность: для каждого значения x_i искать вероятность того, что аппроксимирующая случайная величина попадёт в промежуток $[x_i - 0.5, x_i + 0.5]$.

Ошибка аппроксимации

Чем симметричнее распределение Пуассона, тем точнее его аппроксимирует нормальное. При выполнении условия аппроксимации распределение Пуассона достаточно симметрично, и ошибка аппроксимации мала.