Python задание 5: моделирование

1 Генерация распределений

Давайте разберемся, как сгенерировать дискретные и абсолютно непрерывные случайные величины, используя равномерно распределенные точки на отрезке $[0;1]$.

Для начала кратко напомним генерацию дискретных случайных величин. Предположим, у вас есть случайная величина X с тремя возможными исходами $\{a,b,c\}$ и вероятностями $\{p_a,p_b,p_c\}$ соответственно, где $p_a+p_b+p_c=1$.

1. **Разделение отрезка $[0;1]$:** разделите отрезок $[0;1]$ на три части, где длины интервалов соответствуют вероятностям p_a,p_b,p_c. Например, первый интервал будет $[0,p_a]$, второй $(p_a,p_a+p_b]$, третий $(p_a+p_b,1]$.

2. **Генерация точки:** Сгенерируйте равномерно распределенную случайную точку Y в интервале $[0;1]$.

3. **Определение значения:** Если Y попадает в первый интервал, присвойте X значение a, если во второй — b, если в третий — c.

Эта простая схема легко может быть обобщена на любое распределение дискретной случайной величины.

Генерация абсолютно непрерывных случайных величин может быть выполнена с помощью квантильного преобразования. Здесь вам понадобится обратная функция к функции распределения F_X вашей непрерывной случайной величины X.

1. **Обратная функция распределения:** Убедитесь, что функция распределения F_X непрерывна и строго монотонна, что позволяет определить её обратную функцию F_X^{-1}.

2. **Генерация точки:** Сгенерируйте равномерно распределенную случайную точку Y в интервале $[0;1]$.

3. **Квантильное преобразование:** Значение вашей случайной величины X будет равно $F_X^{-1}(Y)$, то есть вы применяете обратную функцию распределения к сгенерированной точке.

1.1 Задание № 1

Напишите функциидля генерации случайных величин имеющих следующие распределения:

1. Бернулли с параметром p;
2. Биномиальное с параметрами n, p;
3. Геометрическое с параметром p;
4. Пуассона с параметром λ;
5. Равномерное на отрезке $[a; b]$;
6. Показательное с параметром α;
7. Лапласа с параметром α;
8. Нормальное с параметрами a и σ^2 (если Y — стандартное нормальное распределение, то в этом пункте речь идёт о $\sigma Y + a$);
9. Распределения Коши;
10. Распределения, заданного ядром $f(t) = \frac{1}{t^3} I_{\{t>1\}}$.

Продемонстрируйте их работу построив гистограммы для выборок из этих распределений. Сравните эти гистограммы со столбчатыми диаграммами соответствующих дискретных распределений и плотностями для абсолютно непрерывных распределений.

2 Выборочные характеристики

Математическое ожидание и выборочное среднее.

Математическое ожидание $\mathbb{E}X$ представляет собой среднее значение случайной величины в теоретическом смысле, ожидаемое при бесконечном числе наблюдений. Выборочное среднее $\bar{X}$ — это среднее значение, вычисленное на основе конкретной выборки данных.

Выборочное среднее $\bar{X}$ служит оценкой математического ожидания $\mathbb{E}X$. При увеличении объема выборки, выборочное среднее стремится к истинному значению математического ожидания.

Дисперсия и выборочная дисперсия.

Дисперсия $\mathbb{D}X$ измеряет степень разброса значений случайной величины вокруг её математического ожидания. Выборочная дисперсия S^2 оценивает этот разброс на основе ограниченной выборки данных.

Выборочная дисперсия S^2 является оценкой дисперсии $\mathbb{D}X$. С увеличением объема выборки точность этой оценки возрастает.

Длинные и широкие хвосты.

Хвостом распределения называют область больших отклонений от среднего значения. Важна скорость убывания вероятности:

- *Длинный хвост* означает медленное убывание плотности при больших $|x|$, так что большие отклонения встречаются не настолько редко.
- *Короткий хвост* означает быстрое убывание плотности и редкость больших отклонений.
- *Широкий хвост* означает, что заметная часть вероятности находится далеко от центра.

Например, у нормального распределения хвосты короткие (e^{-x^2} убывает очень быстро), а у распределения Коши — длинные и широкие ($1/x^2$ убывает медленно), поэтому большие отклонения встречаются значительно чаще. Не смотря на то, что оба распределения могут принимать любые значения из $\mathbb{R}$.

Асимметрия и выборочная асимметрия.

Коэффициент асимметрии случайной величины X:

$$\gamma_1 = \frac{\mathbb{E}[(X - \mathbb{E}X)^3]}{(\mathbb{D}X)^{3/2}}.$$

Выборочный коэффициент асимметрии определяется для выборки $X_1, \ldots, X_n$:

$$\hat{\gamma}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{S^3}.$$

Коэффициент асимметрии характеризует степень и направление асимметрии распределения относительно его среднего значения.

Для одномодального распределения (с одним пиком) отрицательная асимметрия обычно указывает на то, что хвост распределения расположен слева, а положительная — справа. Однако, если один хвост длинный, а другой широкий, асимметрия не подчиняется простому правилу. Например, нулевое значение асимметрии означает, что хвосты по обе стороны от среднего в целом уравновешены; это характерно для симметричного распределения, но может быть и в асимметричном распределении с одним длинным и тонким хвостом и другим коротким, но широким. Поэтому оценивать симметричность распределения, опираясь только на коэффициент асимметрии, рискованно.

Выборочный коэффициент асимметрии $gam\bar{m}a_1$ используется как оценка истинного коэффициента асимметрии.

Эксцесс и выборочный эксцесс.

Коэффициент эксцесса случайной величины X:

$$\gamma_2 = \frac{\mathbb{E}[(X - \mu)^4]}{(\mathbb{D}X)^2} - 3.$$

Выборочный коэффициент эксцесса определяется для выборки $X_1, \ldots, X_n$:

$$\hat{\gamma_2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{S^4} - 3.$$

Эксцесс, обычно сравниваемый со значением 0, характеризует "тяжеловесность" хвостов распределения. У нормального распределения эксцесс равен 0 (именно для этого и вычитается −3). Отрицательный эксцесс указывает на платикуртическое распределение, которое не обязательно имеет плоскую вершину, но характеризуется меньшим количеством экстремальных выбросов по сравнению с нормальным распределением. Например, равномерное распределение является платикуртическим. Положительный эксцесс свидетельствует о лептокуртическом распределении. Распределение Лапласа, например, имеет хвосты, которые убывают медленнее, чем у нормального распределения, что приводит к большему числу выбросов.

Чтобы упростить сравнение с нормальным распределением, эксцесс рассчитывается как куртозис Пирсона минус 3. Некоторые авторы и программные пакеты используют термин "куртозис" для обозначения именно эксцесса, но здесь для ясности они различаются.

Выборочный коэффициент эксцесса $\hat{\gamma}_2$ служит оценкой истинного коэффициента эксцесса γ_2.

Чтобы лучше понять влияние асимметрии и эксцесса на форму распределения, рассмотрим несколько графиков.

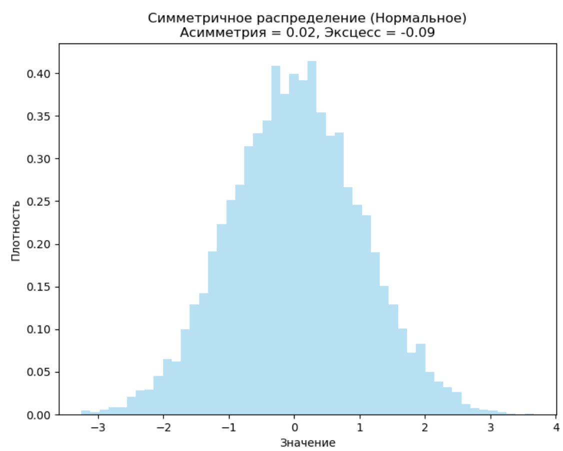

Рис. 1: Симметричное нормальное распределение

На рисунке 1 представлено симметричное нормальное распределение с асимметрией и эксцессом, близкими к нулю. Это указывает на то, что распределение имеет схожие хвосты и форму колокола.

Рисунок 2 показывает распределение с положительной асимметрией. Хвост распределения удлинен вправо, что указывает на присутствие экстремальных значений выше среднего.

На рисунке 3 представлено распределение с отрицательной асимметрией. Здесь хвост удлинен влево, свидетельствуя о наличии экстремальных значений ниже среднего.

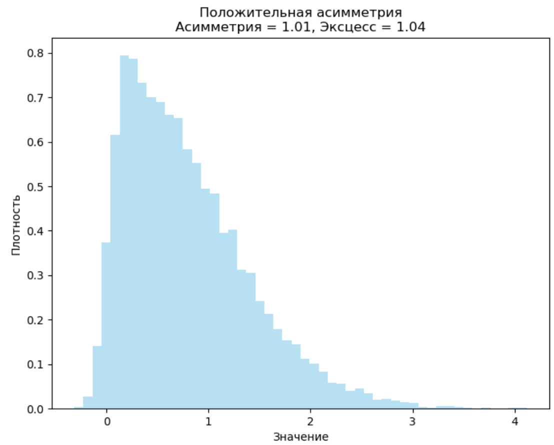

Рис. 2: Распределение с положительной асимметрией

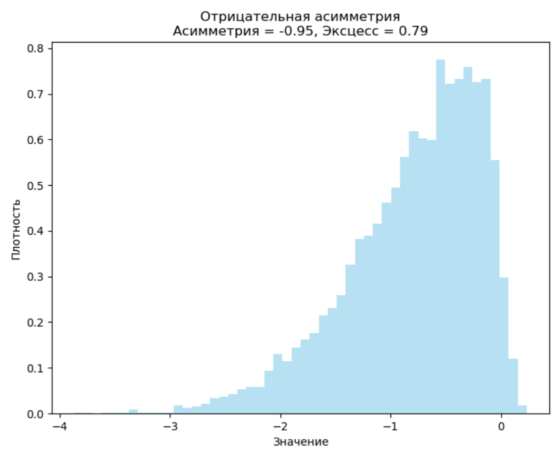

Рис. 3: Распределение с отрицательной асимметрией

Рисунок 4 иллюстрирует лептокуртическое распределение с положительным эксцессом. Такое распределение характеризуется острой вершиной и “тяжелыми” хвостами, что означает повышенную вероятность экстремальных значений.

На рисунке 5 показано платикуртическое распределение с отрицательным эксцессом. Такое распределение имеет более плоскую вершину и “легкие” хвосты, указывая на меньшую вероятность экстремальных значений по сравнению с нормальным распределением.

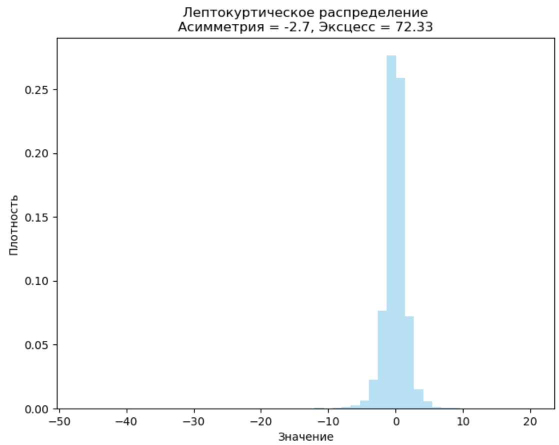

Рис. 4: Лептокуртическое распределение (положительный эксцесс)

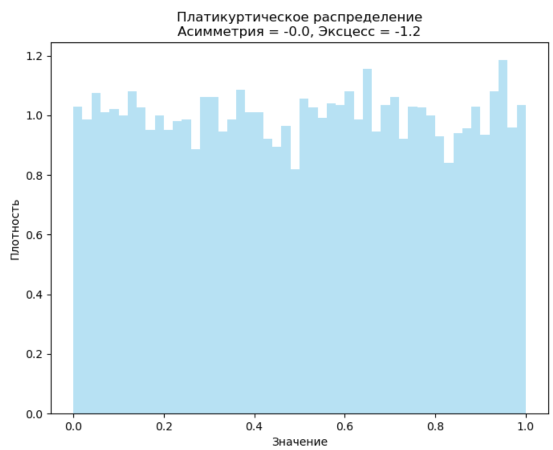

Рис. 5: Платикуртическое распределение (отрицательный эксцесс)

2.1 Задание №2

Исследуйте выборочные асимметрию и эксцесс распределений из первого задания при разных параметрах. Также исследуйте их для гамма и бета распределений. Здесь можно использовать встроенные функции для генерации случайных чисел.

3 Центральная предельная теорема

Если $\xi_1, \xi_2, \ldots, \xi_n$ — независимые и одинаково распределённые случайные величины с ненулевой конечной дисперсией $0 < \mathbb{D}\xi_1 < \infty$, то имеет место следующее приближение:

$$\mathbb{P}\left(\frac{\xi_1 + \ldots + \xi_n - n\mathbb{E}\xi_1}{\sqrt{n\mathbb{D}\xi_1}} \in A\right) \approx \mathbb{P}(\eta \in A),$$

где η имеет стандартное нормальное распределение.

Для нас это даёт способ приблизительной генерации нормального распределения из любого другого: нужно лишь посчитать стандартизацию суммы случайных величин $\frac{\xi_1 + \ldots + \xi_n - n\mathbb{E}\xi_1}{\sqrt{n\mathbb{D}\xi_1}}$.

3.1 Задание №3

Проверьте что центральная предельная теорема работает. Для этого сгенерируйте несколько случайных величин, имеющих стандартное нормальное распределение по только что описанной схеме. Для генерации каждой нормальной случайной величины понадобится своя последовательность $\xi_1, \xi_2, \ldots, \xi_n$:

$$\xi_1^{(1)}, \xi_2^{(1)}, \ldots, \xi_n^{(1)} \to \eta_1;$$
$$\xi_1^{(2)}, \xi_2^{(2)}, \ldots, \xi_n^{(2)} \to \eta_2;$$
$$\ldots$$
$$\xi_1^{(N)}, \xi_2^{(N)}, \ldots, \xi_n^{(N)} \to \eta_N.$$

Сделайте возможность выбирать n и N. Нарисуйте на одном графике гистограмму по значениям $\eta_1, \eta_2, \ldots, \eta_N$ и плотность стандартного нормального распределения для сравнения.