Практическая работа № 3. Компьютерное моделирование выборок непрерывных случайных величин, первичная обработка выборки

3.1. О содержании и задачах практической работы

В настоящей практической работе обсуждаются три задачи:

1. Компьютерное моделирование выборок непрерывных случайных величин с заданным законом распределения.

2. Первичная обработка выборки для визуализации закона распределения генеральной совокупности.

3. Иллюстрация фундаментальных законов теории вероятностей (статистическая «проверка» основных теорем).

О первой задаче. Существует несколько методов моделирования выборок непрерывных случайных величин. Одним из них является *метод обратных функций*. Он основан на следующем, легко проверяемом утверждении. Если X – непрерывная случайная величина с функцией распределения $F_X(X)$, то случайная величина $Y = F_X(X)$ равномерно распределена на отрезке $[0,1]$, т.е. $Y \sim R(0,1)$. Следовательно, можно рассматривать случайную величину X как корень уравнения $Y = F_X(X)$, т.е. $X = F_X^{-1}(Y)$. Опираясь на это представление, действуют так: моделируют выборку случайной величины Y, равномерно распределенной на отрезке $[0,1]$, а затем по формуле $X = F_X^{-1}(Y)$ рассчитывают выборочные значения случайной величины X.

Например, если $X \sim Ex(\lambda)$, то при $x \geq 0$ $F_X(x) = 1 - e^{-\lambda x}$, т.е. $1 - e^{-\lambda X} = Y$, откуда $X = -\frac{1}{\lambda}\ln(1-Y)$. Заметим, что можно использовать формулу попроще. Действительно, так как $Y \sim R(0,1)$ то и

$(1-Y) \sim R(0,1)$. Поэтому вместо формулы $X = -\frac{1}{\lambda}\ln(1-Y)$ для моделирования можно использовать формулу $X = -\frac{1}{\lambda}\ln Z$, где $Z \sim R(0,1)$.

Применяют и другие методы моделирования: *метод исключения* (*метод отбора*), всевозможные специальные представления для конкретных распределений (см. например, [2, с. 193]).

В языке программирования Python широко представлены различные средства моделирования выборок (реализующие, в том числе, и упомянутые методы), в разделе 3.3 настоящего практикума приведены некоторые из них.

О второй задаче. Часто требуется по результатам экспериментов – выборке значений непрерывной случайной величины составить представление о законе распределения (функции или плотности распределения). И первым шагом в этом направлении является построение гистограммы относительных частот – статистического аналога плотности распределения. Однако насколько успешно гистограмма справится с задачей приближенного описания закона распределения, напрямую зависит от способа и числа интервалов группировки. Распространенный подход – использование интервалов одной длины (дальше мы будем обсуждать именно этот случай). Но какое число интервалов лучше использовать? Исходя из общих соображений, число интервалов следует выбирать так, чтобы вид гистограммы был «как можно ближе» к графику плотности распределения генеральной совокупности. Естественно предположить, что «оптимальное» число интервалов будет зависеть и от выбранной «меры близости», и от вида закона распределения генеральной совокупности.

О третьей задаче. В фокусе внимания – центральная предельная теорема. Будем работать с одним из ее вариантов (в формулировке Ляпунова для одинаково распределенных слагаемых).

Если случайные величины в последовательности $X_1, X_2, ..., X_n, ...$ *независимы, одинаково распределены и имеют конечные математиче-*

ское ожидание m и дисперсию σ^2, *то для любого действительного* x

$$\lim_{n\to\infty} P\left\{\frac{\sum_{k=1}^{n} X_k - nm}{\sigma\sqrt{n}} < x\right\} = \Phi(x).$$

Если позволить себе нестрогость, то суть центральной предельной теоремы можно выразить следующим образом: какой бы закон распределения не имели независимые случайные величины, стандартизированная сумма большого числа таких величин с вероятностью, близкой к единице, является стандартизированной нормальной величиной. Или еще, более вольно: сумма большого числа случайных величин при весьма широких условиях распределена «почти» нормально – при больших n можно считать, что $\sum_{k=1}^{n} X_k \sim N\left(nm, \sigma\sqrt{n}\right)$.

Компьютерное моделирование выборок позволяет наглядно проиллюстрировать это фундаментальное теоретическое утверждение. Сгенерируем «много» выборок «большого» объема одной и той же генеральной совокупности и поэлементно сложим их. По полученной таким образом выборке построим гистограмму. Центральная предельная теорема «предсказывает» нам «близость» гистограммы и графика плотности нормального распределения. Наложим график на гистограмму и убедимся в соответствии полученного результата утверждению центральной предельной теоремы.

3.2. Математические понятия и утверждения: краткая информация и ссылки на источники

1. Центральная предельная теорема в формулировке Ляпунова для одинаково распределенных слагаемых [1, с. 127 – 131; 2, с. 134 – 136].

2. Моделирование случайных величин методом обратных функций [2, с. 192 – 193].

3. Генеральная совокупность; случайная выборка объема n; эмпирическая выборка (выборка) объема n; эмпирическая (выборочная) функция распределения, теорема Гливенко; выборочное математическое ожидание, выборочная дисперсия, исправленная выборочная дисперсия, квартили [1, с. 132 – 136; 2, с. 185 – 191].

4. Группированный статистический ряд; гистограмма и полигон относительных частот [1, с. 132 – 136; 2, с. 185 – 191].

Большинство рекомендуемых формул для оценки числа k интервалов группировки носит эмпирический характер. Приведем некоторые *правила выбора числа интервалов группировки* (все они реализованы в библиотечных инструментах языка программирования Python). Далее, n – объем выборки, h – длина каждого из интервалов.

Правило квадратного корня: $k = \sqrt{n}$.

Правило Райса: $k = 2\sqrt[3]{n}$.

Правило Стерджесса: $k = \left[1 + \log_2 n\right]$.

Формула Скотта: $k = \left[\frac{x_{\max} - x_{\min}}{h}\right]$, $h = \sqrt[3]{\frac{24\sqrt{\pi}}{n}}s$ ($x_{\min}$ и $x_{\max}$ – наименьший и наибольший элементы выборки соответственно).

Правило Фридмана – Диакониса: $k = \mathrm{ceil}\left(\frac{x_{\max} - x_{\min}}{h}\right)$, $h = \frac{2 \cdot \mathrm{IQR}}{\sqrt[3]{n}}$ ($\mathrm{ceil}(a)$ – ближайшее целое, большее или равное a, $\mathrm{IQR} = Q_3 - Q_1$ – разность между третьей и первой квартилью выборки).

Наша задача – на качественном уровне (визуально) оценить влияние числа интервалов группировки на близость гистограммы относительных частот к теоретической плотности распределения генеральной совокупности.

3.3. Библиотечные инструменты языка программирования Python

Загрузка основных модулей

```
import numpy as np
import scipy.stats as sts
import scipy.special as sc
import matplotlib.pyplot as plt
%matplotlib inline
```

Генераторы выборок некоторых непрерывных распределений в библиотеке numpy

Функция `np.random.normal(loc, scale, size)`
Возвращает выборку заданного объема `size` (если `size` – число) нормального распределения $N(m,\sigma)$.

Параметры: `loc`= m ; `scale`= σ . Если `size` – кортеж, то генерируется массив заданной формы.

Функция `np.random.uniform(low=0.0, high=1.0, size=None)`
Возвращает выборку заданного объема `size` (если `size` – число) равномерного распределения $R(a,b)$.

Параметры: `low`= a ; `high`= b . Если `size` – кортеж, то генерируется массив заданной формы.

Функция `np.random.exponential (scale=1.0, size=None)`
Возвращает выборку заданного объема `size` (если `size` – число) экспоненциального распределения $Ex(\lambda)$.

Параметр: `scale` = $1/\lambda$ (равен математическому ожиданию). Если `size` – кортеж, то генерируется массив заданной формы.

Функция `np.random.chisquare (df, size=None)`
Возвращает выборку заданного объема `size` (если `size` – число) распределения хи-квадрат с `df` степенями свободы. Если `size` – кортеж, то генерируется массив заданной формы.

Функция `np.random.f(dfnum, dfden, size)`
Возвращает выборку заданного объема `size` (если `size` – число) распределения Фишера со степенями свободы `dfnum`, `dfden`. Если `size` – кортеж, то генерируется массив заданной формы.
В модуле numpy.random также имеются генераторы выборок следующих распределений: бета, гамма, Гумбеля, Лапласа, логистического, логнормального, степенного, Рэлея, треугольного, Ломакса (Парето II вида), фон Мизеса, Уайльда, Вейбулла и др.

Построение эмпирической функции распределения в модуле `statsmodels.distributions.empirical_distribution`

Функция `ECDF`
`statsmdels.distributions.empirical_distribution.ECDF(x,side=right)`
Возвращает эмпирическую функцию распределения.
Параметры: `x` – массив (выборка); `side` – задает форму интервалов, по которым строятся ступени эмпирической функции: `right` (по умолчанию) – интервалы вида $[...)$, открытые справа, `left` – интервалы вида $(...]$, открытые слева.

Средства визуализации: построение гистограммы и эмпирической функции распределения в модуле matplotlib.pyplot

Функция `plt.hist(x, bins=None, density=None, weights=None, cumulative=False, histtype='bar', align='mid', orientation='vertical', log=False, color=None)`
Строит гистограмму и возвращает два массива: высот столбцов гистограммы и центров интервалов группировки.
Параметры: `x` – массив (выборка) ; `bins` – число интервалов группировки или последовательность, задающая границы интервалов (все интервалы, кроме последнего, полуоткрытые вида $[...)$), или строка из списка, который приводится после перечня параметров; `density` – если `True`, то строится гистограмма относительных частот (суммарная площадь прямоугольников равна 1); `weights` – массив весов той же формы, что и `x`; `cumulative` – если `True`, то в сочетании с признаком `density=True` строит эмпирическую функцию распределения; `histtype` – кроме типа `'bar'` можно указать `'barstacked'` и `'step'`; `align` – задает расположение центров прямоугольников; `orientation` – установив значение `'horizontal'`, можно повернуть график на 90°; `log` – если `True`, для осей используется логарифмическая шкала `color` – признак, устанавливающий цвет.
По умолчанию число интервалов группировки равно 10.

Список правил для выбора числа интервалов:
`bins='auto'` – максимальное из значений, получаемых по правилу Стерджесса и Фридмана – Диаконисса;
`bins='fd'` – правило Фридмана – Диаконисса;
`bins=' sturges'` – правило Стерджесса;
`bins= 'doane'` – правило Доэна;
`bins=' scott'` – правило Скотта;
`bins='stone'` – обобщение правила Скотта;
`bins=' rice'` – правило Райса;
`bins=' sqrt'` – правило квадратного корня.

3.4. Примеры для совместного обсуждения

Пример 1. Рассмотрим вопросы, связанные с первичной обработкой выборки. В качестве «опытного» материала будем использовать выборки значений равномерно распределенной на отрезке $[a,b]$ случайной величины X, генерируя их с помощью библиотечных инструментов языка программирования Python (в приложении к практической работе № 3 рассматривается случайная величина $X \sim R(2,6)$). Будем работать по следующему плану:

1) Сгенерируем выборку объема 50. Построим график эмпирической функции распределения и теоретической функции распределения (в одной системе координат). Повторим действия с выборками объемов 100, 500, 1000. Какие закономерности можно подметить? Иллюстрацией какого теоретического утверждения служат полученные рисунки?

2) Для выборки объема 50 для 10 интервалов группировки построим гистограммы относительных частот и график теоретической плотности распределения (в одной системе координат). Повторим действия с выборками объемов 100, 500, 1000. Какие закономерности можно подметить? Какое теоретическое утверждение поясняют полученные рисунки?

3) Убедимся, что способы построения группированной выборки влияют на качество оценки теоретической плотности распределения с помощью гистограммы относительных частот. С этой целью сгенерируем выборку объема 500 и найдем для нее ряды распределения при 5, 10,

15, 25 интервалах группировки. В каждом случае оценим точность представления теоретической плотности распределения соответствующим рядом. В качестве показателя оценки можно взять максимальное на отрезке $[a,b]$ значение модуля разности гистограммы и теоретической плотности.

Фрагмент выполнения этого примера приведен в приложении к практической работе № 3.

Пример 2. Сгенерируйте выборку объема 100 нормально распределенной случайной величины $X \sim N(m,\sigma)$ (параметры m,σ выберите самостоятельно).

1) В одной системе координат постройте графики теоретической и эмпирической функций распределения.

2) Постройте гистограммы, последовательно используя различные правила определения числа интервалов (квадратного корня, Райса, Стерджесса, Скотта и Фридмана – Диакониса). Для каждого случая оцените визуально качество представления теоретической плотности распределения гистограммой.

Проделайте задания 1) и 2) для выборки объема 1000.

Сопоставьте ваши наблюдения для двух выборок.

3.5. Задания для самостоятельного выполнения

Задание 1. 1) Используя метод обратных функций, сформируйте выборку объёма N = (100 + длина вашей фамилии) случайной величины X, распределенной по показательному закону с параметром $\lambda = 0{,}1 \cdot$ (длину вашего имени). В одной системе координат постройте график эмпирической функции распределения, а также теоретической функции распределения случайной величины X. Прокомментируйте полученные рисунки.

2) Рассчитайте число интервалов группировки по различным правилам (квадратного корня, Райса, Стерджесса, Скотта и Фридмана – Диакониса). Для каждого случая оцените качество представления теоретической плотности распределения гистограммой (в качестве показателя оценки возьмите максимальное по модулю отклонение гистограммы

от теоретической плотности). Для какого правила получился лучший по точности результат? Расставьте правила «по рейтингу» точности. Сгенерируйте несколько новых выборок и повторите расчеты. Можно ли сказать, что рейтинг устойчив?

Задание 2. Задание выполняется по вариантам. Пусть $X_1, X_2, X_3, \ldots$ – независимые случайные величины, распределенные по тому же закону, что и случайная величина X из вашего варианта.

1) Постройте гистограммы по выборкам объема 200 для следующих случайных величин: X, $\frac{1}{5}\sum_{i=1}^{5} X_i$, $\frac{1}{50}\sum_{i=1}^{50} X_i$, $\frac{1}{500}\sum_{i=1}^{500} X_i$.

2) К какому теоретическому распределению приближается случайная величина из $\frac{1}{500}\sum_{i=1}^{500} X_i$? На каком основании можно строится ваше предположение? Проиллюстрируйте ваше заключение, наложив на гистограмму график плотности этого распределения.

Вариант 1. X имеет распределение Лапласа с параметрами $\alpha = 1{,}5,\ \beta = 0$.

Вариант 2. X имеет распределение Фишера с параметрами 6, 58.

Вариант 3. X имеет распределение хи-квадрат с 4 степенями свободы.

Вариант 4. X имеет гамма-распределение с параметрами 1 и 2.

Вариант 5. X имеет бета-распределение с параметрами 2 и 5.

Вариант 6. X имеет распределение Лапласа с параметрами $\alpha = 0{,}4,\ \beta = 0$.

Вариант 7. X имеет распределение Фишера с параметрами 6, 20.

Вариант 8. X имеет логнормальное распределение с параметрами 0 и 1,4.

Вариант 9. X имеет логнормальное распределение с параметрами 1 и 1,2.

Вариант 10. X имеет распределение Лапласа с параметрами $\alpha = 2,\ \beta = 1$.

Вариант 11. X имеет распределение хи-квадрат с 3 степенями свободы.

Вариант 12. X имеет бета-распределение с параметрами, равными 1 и 1.

Вариант 13. X имеет распределение Фишера с параметрами 9, 10.

Вариант 14. X имеет распределение хи-квадрат с 5 степенями свободы.