# Практическая работа № 4.
# Точечное оценивание параметров распределения по выборке

## 4.1. О содержании и задачах практической работы

В настоящей практической работе обсуждаются три задачи.

1. Поиск точечных оценок числовых параметров генеральной совокупности по выборке с использованием стандартных библиотечных средств языка программирования Python.

2. Изучение свойств точечных оценок параметров распределения, а также сравнительный анализ различных точечных оценок одного параметра с использованием сгенерированных выборок генеральной совокупности.

3. Поиск точечных оценок параметров распределения по результатам компьютерного моделирования статистического эксперимента.

*О первой задаче.* В языке программирования Python имеется много инструментов для получения точечных оценок основных числовых характеристик распределений. Наша задача – научиться ими пользоваться.

*О второй задаче*. Безусловно, основные точечные оценки хорошо изучены теоретически. Какие-то результаты мы можем получить аналитически и сами. Поэтому задача состоит не в том, чтобы открыть новые факты, а в том, чтобы усвоить известные понятия и утверждения. Попробуем поработать над пониманием свойств точечных оценок, используя наше умение моделировать любое число выборок большого объема с помощью программных средств. Имея выборки, мы можем визуализировать свойства точечных оценок, наглядно их иллюстрировать и тем самым понять связанные с их использованием возможности и ограничения.

*О третьей задаче*. Положим, проводится статистический эксперимент. В этом эксперименте наблюдается случайная величина, и нас интересуют ее числовые характеристики. Чтобы их найти, вначале можно получить путем рассуждений закон распределения этой случайной величины, а затем, используя его, вычислить математическое ожидание, дисперсию и т.д. Однако аналитический поиск закона распределения и числовых характеристик может оказаться сложной и трудоемкой задачей. Допустим, мы готовы довольствоваться оценками числовых характеристик (начальных и центральных моментов). В таком случае можно разработать компьютерную модель статистического эксперимента, провести компьютерный эксперимент большое число раз, получить выборку и по ней оценить нужные нам числовые характеристики распределения.

## 4.2. Математические понятия и утверждения: краткая информация и ссылки на источники

**1.** Точечные оценки параметров распределения. Характеристики качества оценок: несмещенность (в том числе асимптотическая), состоятельность, эффективность. Свойства выборочного среднего, выборочной дисперсии, исправленной выборочной дисперсии, начальных и центральных моментов [1, с. 151 – 161; 2, с. 218 – 225].

**2.** Визуализация числовых характеристик выборки с боксплотов.

Боксплот (BoxPlot, «ящик с усами») – вид диаграммы, компактно изображающей одномерное распределение данных. Показывает медиану, нижний и верхний квартили, минимальное и максимальное значения выборки и выбросы. Расстояния между различными частями боксплота позволяют определить степень разброса и асимметрию данных (рис.1).

Границами боксплота служат первый и третий квартили, линия в середине боксплота – медиана. Концы «усов» – края статистически значимой выборки (без выбросов), и они могут определяться поразному. Стандартный подход: левый край находят по формуле $Q_3 - 1{,}5(Q_3 - Q_1)$,

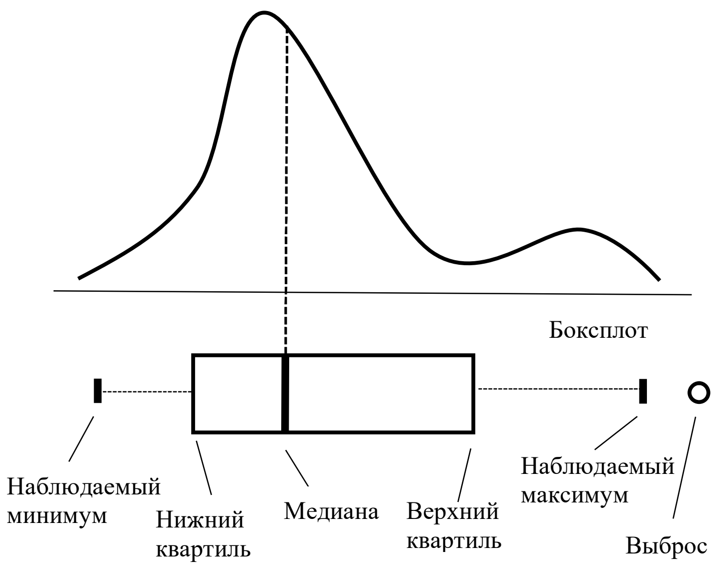


*Рис.1.*

## 4.3. Библиотечные инструменты языка программирования Python

**Загрузка основных модулей**

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

**Точечные оценки параметров распределения в пакете numpy**

Функция `np.mean(a, axis)`
Возвращает выборочное среднее.
Параметры: `a` – массив; в случае многомерного массива `a` можно указать ось (`axis`), вдоль которой вычисляется среднее.
Функция `np.nan mean(a, axis)` при вычислении игнорирует пропущенные данные `nan` (важно при обработке реальных данных).
Для вычисления выборочных начальных моментов порядка $k$ можно использовать функцию `mean` применительно к $k$-й степени массива `a`.

Функция `np.var(a, axis, ddof)`
Возвращает оценку дисперсии по выборке `a`.
Параметры: `a` – массив; в случае многомерного массива `a` можно указать ось (`axis`), вдоль которой вычисляется дисперсия; `ddof` по умолчанию равен 0 (вычисляется выборочная дисперсия), если задать `ddof=1`, то функция возвращает исправленную выборочную дисперсию.
Функция `np.nanvar(a, axis, ddof)` при вычислении игнорирует пропущенные данные `nan`.

Функция `np.std(a, axis, ddof)`
Возвращает корень из выборочной (или исправленной выборочной) дисперсии.
Параметры: `a` – массив; в случае многомерного массива `a` можно указать ось, вдоль которой вычисляется дисперсия; `ddof` по умолчанию равен 0 (вычисляется выборочная дисперсия), если задать `ddof=1`, то функция возвращает исправленную выборочную дисперсию.

Функция `np.median(a, axis=None, out=None)`
Возвращает выборочную медиану (вычисляется как центральный элемент $a_{\frac{n-1}{2}}$ отсортированного по неубыванию массива `a` при нечетном $n$ и как среднее арифметическое двух центральных значений при четном $n$).

Параметры: `a` – массив; в случае многомерного массива `a` можно указать ось (`axis`), вдоль которой вычисляется среднее; `out` – массив, если он указан, в него помещаются вычисленные значения медиан.

Функция `np.quantile (a, q, axis=None, out=None, interpolation='linear')`
возвращает квантиль порядка `q` (указывается число из интервала $(0,1)$).

Параметры: `a` – массив; в случае многомерного массива `a` параметр `axis` – ось (кортеж осей), вдоль которой производятся вычисления; `out` – массив, если он указан, в него помещаются вычисленные значения квантилей; `interpolation` – признак, определяющий метод интерполяции в ситуации, когда квантиль расположена между двумя значениями массива (`'linear'` по умолчанию, есть другие варианты).

*Первый квартиль* $Q_1$ выборки $X$ вычисляется с помощью функции `np.quantile(X,0.25)`.

*Вторым квартилем* $Q_2$ выборки $x$ называется квантиль порядка 0,5 (медиана).

*Третий квартиль* $Q_3$ выборки $X$ вычисляется с помощью функции `np.quantile(X,0.75)`.

**Точечные оценки параметров распределения в модуле scipy.stasts**

Функция `sts.moment(x, moment=k, axis=0, nan_policy='propogate')`
Возвращает выборочный центральный момент порядка $k$.

Параметры: `x` – выборка; `axis` – ось, вдоль которой вычисляется оценка; `nan_policy` – определяет способ обработки пропущенных значений (`'propogate'` – возвращает `nan`, `'raise'` – генерирует ошибку, `'omit'` – игнорирует пропущенные данные).

Функция `sts.skew(x, axis=0, bias=True, nan_policy='propogate')`
Возвращает выборочный коэффициент асимметрии.

Параметры: `x` – выборка; `axis` – ось, вдоль которой вычисляется оценка; `bias` – признак (если `False` – применяется коррекция для устранения смещенности); `nan_policy` – определяет способ обработки пропущенных значений.

Функция `sts.kurtosis(x, axis=0, fisher=True, bias=True, nan_policy='propogate')`
Возвращает выборочный коэффициент асимметрии.
Параметры: `x` – выборка; `axis` – ось, вдоль которой вычисляется оценка; `fisher` – признак, если равен `True` (по умолчанию), то в формуле для эксцесса из отношения моментов вычисляется число 3; `bias` – признак (если `False` – применяется коррекция для устранения смещенности); `nan_policy` – определяет способ обработки пропущенных значений.

Функция `sts.igr(x)`
Вычисляет межквартильный размах – разность между третьим и первым квартилями.

Функция `sts.describe(a, axis, ddof, bians, nan_policy)`
Возвращает набор оценок основных параметров случайной величины: `nobs` – объем выборки; `minmax` – кортеж, содержащий максимальное и минимальное значение выборки; `mean` – выборочное среднее; `variance` – исправленная выборочная дисперсия $s^2$ (в случае задания `ddof=1` или по умолчанию) либо выборочная дисперсия (в случае задания `ddof=0`); `skewness` – коэффициент асимметрии; `kurtosis` – коэффициент эксцесса (в случае задания `bias=False`, коэффициенты асимметрии и эксцесса корректируются на величину смещения).
Параметры: `a` – выборка; `axis` – задание оси (для многомерной выборки); `ddof` – признак смещенности (только для дисперсии); `bians` – признак коррекции (только для асимметрии и эксцесса); `nan_policy` – задает способ обработки пропущенных данных.

**Средства визуализации: построение гистограммы и боксплота в пакете seaborn**

Функция `boxplot(x=None, y=None, hue=None, data=None, order=None, hue_order=None, orient=None, color=None)`

Строит боксплот:

Параметры: `x,y, hue` – наименование признаков в наборе `data;` `data` – датафрейм, или массив `numpy`, или список; `order` и `hue_order` – строки, с помощью которых можно изменить порядок вывода признаков на график; `orient` – вертикальная или горизонтальная ориентация («`v`» или «`u`»); `color` – задание цвета.

Построение боксплота в пакете удобно сочетать с построением гистограммы с помощью функции `histplot`.

## 4.4. Примеры для совместного обсуждения

**Пример 1.** Рассмотрим способы нахождения и визуализации точечных оценок параметров распределения по выборке с применением инструментов языка программирования Python. Сгенерируем выборку какого-либо распределения и используем ее в качестве «опытных» данных.

1) Найдем точечные оценки числовых характеристик распределения.

2) Визуализируем распределение данных с помощью гистограммы и боксплота.

3) Поупражняемся в интерпретации («чтении») боксплота. Будем генерировать выборки, меняя параметры распределения, находить по выборкам точечные оценки параметров, строить гистограммы, боксплоты и сопоставлять найденные характеристики.

Вариант выполнения п. 1) и 2) приведен в приложении к практической работе № 4, пример 1.

**Пример 2.** Поставим задачу экспериментально изучить свойства выборочных оценок. Сделаем это на примере выборочного центрального момента четвертого порядка $\mu_4^*\left(x_1, x_2, ..., x_n\right)$ и нормально распределенной генеральной совокупности $X$ с какими-нибудь конкретными параметрами $m, \sigma$.

1) Проиллюстрируем состоятельность оценки. Сгенерируем последовательность выборок возрастающего объема и для каждой из них

рассчитаем $\mu_4^*$. Построим график выборочных центральных моментов, откладывая по оси абсцисс объем выборки, а по оси ординат найденное по соответствующей выборке значение $\mu_4^*$. Ожидаемый результат: выраженная тенденция к сближению значений выборочных центральных моментов четвертого порядка и теоретического момента $\mu_4$ генеральной совокупности (что подтверждает сходимость по вероятности выборочных центральных моментов к теоретическому значению, т.е. состоятельность оценки).

2) Попробуем экспериментально «узнать», является ли $\mu_4^*$ несмещенной оценкой центрального момента четвертого порядка. Для этого по выборке вычислим выборочное математическое ожидание, оценим плотность распределения случайной величины $\mu_4[X]$ с помощью гистограммы, а также визуализируем характеристики распределения $\mu_4[X]$ с помощью боксплота. Уточним: для работы нам нужна не выборка значений генеральной совокупности, а выборка значений случайной величины $\mu_4[X]$. Чтобы получить первый элемент такой выборки, следует сгенерировать выборку объема $n$ генеральной совокупности и вычислить по ней значение $\mu_4^*$ (оно и будет первым элементом выборки случайной величины $\mu_4[X]$). Для получения второго элемента нужно заново сгенерировать выборку объема $n$ генеральной совокупности $X$ и по ней найти новое значение $\mu_4^*$ (второй элемент выборки случайной величины $\mu_4[X]$). После многократного повторения этой операции у нас образуется выборка объема $N$ случайной величины $\mu_4[X]$, по которой мы вычислим выборочное среднее. Осталось интерпретировать полученные результаты и сделать вывод (вариант выполнения приведен в приложении к практической работе № 4, пример 2).

## 4.5. Задания для самостоятельного выполнения

**Задание 1_2025.** Пусть случайная величина $X$ имеет равномерное распределение на отрезке $[0,1]$ и $X_1, X_2, ..., X_n$ – случайная выборка объема $n$ генеральной совокупности $X$. Для математического ожидания генеральной совокупности рассмотрим две оценки: выборочное среднее

$\overline{X} = \dfrac{\sum_{k=1}^{n} X_k}{n}$ и случайную величину $\widehat{m} = \dfrac{X^{(1)} + X^{(2)}}{2}$, где $X^{(1)}$ и $X^{(2)}$ - наименьший и наибольший элемент выборки соответственно.

Проведите вычислительные эксперименты, позволяющие визуализировать и сопоставить свойства этих оценок, результаты проанализируйте, сформулируйте выводы.

1) Путем компьютерного моделирования получите выборки различного объема $n$ (рассмотрите последовательность $n$ от 50 до 10 000 с некоторым шагом). Для каждой выборки вычислите значения обеих оценок. Визуализируйте результаты, построив графики последовательностей реализаций значений оценок. Опираясь на полученные результаты, исследуйте вопрос о состоятельности оценок $\overline{X}$ и $\hat{m}$.

2) Зафиксируйте достаточно большое значение $n$ (например, $10\,000$). Путем компьютерного моделирования сгенерируйте $N$ выборок случайных величин $\overline{X} = \dfrac{\sum_{k=1}^{n} X_k}{n}$ и $\widehat{m} = \dfrac{X^{(1)} + X^{(2)}}{2}$. Опираясь на теорему Чебышёва, исследуйте вопрос о несмещенности оценок $\overline{X}$ и $\hat{m}$ (например, для $\overline{X}$ рассуждаем так: по теореме Чебышёва среднее арифметическое выборочных значений случайной величины $\overline{X}$ сходится по вероятности к математическому ожиданию $\overline{X}$, и, следовательно, при больших $N$ среднее арифметическое выборочных значений случайной величины $\overline{X}$ с вероятностью, близкой к единице, будет «очень близко» к $M\left[\overline{X}\right]$, и, значит, при проверке равенства $M\left[\overline{X}\right] = M\left[X\right]$ можно заменить $M\left[\overline{X}\right]$ средним арифметическим выборочных значений случайной величины $\overline{X}$ ).

3) Зафиксируйте достаточно большое значение $n$. Путем компьютерного моделирования сгенерируйте выборки объема $N$ случайных величин – оценок $\overline{X} = \dfrac{\sum_{k=1}^{n} X_k}{n}$ и $\widehat{m} = \dfrac{X^{(1)} + X^{(2)}}{2}$ ($N$ должно быть велико) и визуализируйте их двумя способами: (1) с помощью бокспло-

тов; (2) с помощью гистограммы. Опираясь на полученные результаты, исследуйте вопрос о сравнительной эффективности оценок $\overline{X}$ и $\hat{m}$ .

4) Теоретически проверьте выводы о свойствах оценок $\overline{X}$ и $\hat{m}$ , к которым вы пришли в ходе выполнения заданий 1)-3) (аналитические выкладки запишите «от руки», фото прикрепите к отчету).

**Задание 2_2025.** Задание выполняется по вариантам.

Вариант 1. Имеется выпуклый семиугольник с вершинами $A_1, A_2, \dots, A_7$. В начальный момент времени в вершине $A_1$ находится частица. С вероятностью $\frac{1}{6}$ частица может перейти в любую из вершин, отличных от $A_1$. На втором шаге частица из новой вершины может сноа перейти в любую другую вершину и т.д. Случайная величина $Y$ – номер того шага, на котором частица первый раз возвратится в вершину $A_1$.

1) Методом статистических испытаний получите выборку значений случайной величины $Y$ ; постройте гистограмму и боксплот.

2) Методом статистических испытаний оцените математическое ожидание, медиану и дисперсию случайной величины $Y$, а также коэффициенты асимметрии и эксцесса.

Вариант 2. Имеется выпуклый семиугольник с вершинами $A_1, A_2, \dots, A_7$. В начальный момент времени в вершине $A_1$ находится частица. С вероятностью $\frac{1}{6}$ частица может перейти в любую из вершин, отличных от $A_1$. На втором шаге частица из новой вершины может снова перейти в любую другую вершину и т.д. Случайная величина $Y$ – номер того шага, на котором впервые будут проведены все диагонали многоугольника.

1) Методом статистических испытаний получите выборку значений случайной величины $Y$ ; постройте гистограмму и боксплот.

2) Методом статистических испытаний оцените математическое ожидание, медиану и дисперсию случайной величины $Y$, а также коэффициенты асимметрии и эксцесса.

Вариант 3. Имеется выпуклый шестиугольник с вершинами $A_1, A_2, \dots, A_6$. В начальный момент времени в вершине $A_1$ находится частица. С вероятностью $\frac{1}{5}$ частица может перейти в любую из вершин, отличных от $A_1$. На втором шаге частица из новой вершины может снова перейти в любую другую вершину и т.д. Случайная величина $Y$ –

номер того шага, на котором впервые будут проведены все главные диагонали многоугольника.

1) Методом статистических испытаний получите выборку значений случайной величины $Y$; постройте гистограмму и боксплот.

2) Методом статистических испытаний оцените математическое ожидание, медиану и дисперсию случайной величины $Y$, а также коэффициенты асимметрии и эксцесса.

Вариант 4. На окружности радиуса 8 находится точка. В последовательные моменты времени $i$ $(i = 1, 2, 3, \ldots)$ точка передвигается по окружности (в одну и ту же сторону) на случайное расстояние $X_i$, распределенное по показательному закону с параметром 0,2. Случайная величина $Y$ – номер того шага, на котором точка пересечет свою исходную позицию.

1) Методом статистических испытаний получите выборку значений случайной величины $Y$; постройте гистограмму и боксплот.

2) Методом статистических испытаний оцените математическое ожидание, медиану и дисперсию случайной величины $Y$, а также коэффициенты асимметрии и эксцесса.

Вариант 5. На окружности радиуса 4 находится точка. В последовательные моменты времени $i$ $(i = 1, 2, 3, \ldots)$ точка передвигается по окружности (в одну и ту же сторону) на случайное расстояние $X_i$, распределенное равномерно на отрезке $[1, 3]$. Случайная величина $Y$ – номер того шага, на котором точка пересечет свою исходную позицию.

1) Методом статистических испытаний получите выборку значений случайной величины $Y$; постройте гистограмму и боксплот.

2) 2) Методом статистических испытаний оцените математическое ожидание, медиану и дисперсию случайной величины $Y$, а также коэффициенты асимметрии и эксцесса.