Федеральное государственное автономное образовательное учреждение
высшего образования
**САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО**
Факультет систем управления и робототехники

Отчёт по лабораторной работе №1
по дисциплине «Математическая статистика»

Вариант 3

Студент: Сайфуллин Д.Р.
Поток: Мат Стат 22
Преподаватель: Шкваренко А.А.

Санкт-Петербург
2025 г.

Задание 1

Для выполнения данного задания необходимо выбрать непрерывное распределение, у которого существуют первые четыре момента. Далее требуется экспериментально убедиться в асимптотической нормальности выборочного среднего, выборочной дисперсии и выборочной медианы (квантиль порядка 0.5). Также нужно проверить выполнение:

$$n\,F\big(X_{(2)}\big) \;\to U_1 \sim \Gamma(2,1), \quad n\Big(1 - F\big(X_{(n)}\big)\Big) \;\to U_2 \sim \Gamma(1,1) = \mathrm{Exp}(1).$$

где $X_{(2)}$ — вторая порядковая статистика, $X_{(n)}$ — максимальная порядковая статистика, а F — функция распределения выбранного закона.

В данной работе возьмем распределение $\Gamma(\alpha = 3, \theta = 2)$. Далее $M = 1000$ раз сгенерируем выборки объемом $n = 1000$. Для каждой выборки рассчитаем выборочное среднее, выборочную дисперсию, медиану, найдем второй элемент в выборке, а также максимальное значение. Построим гистограммы и сравним их с теоретическими значениями распределения.

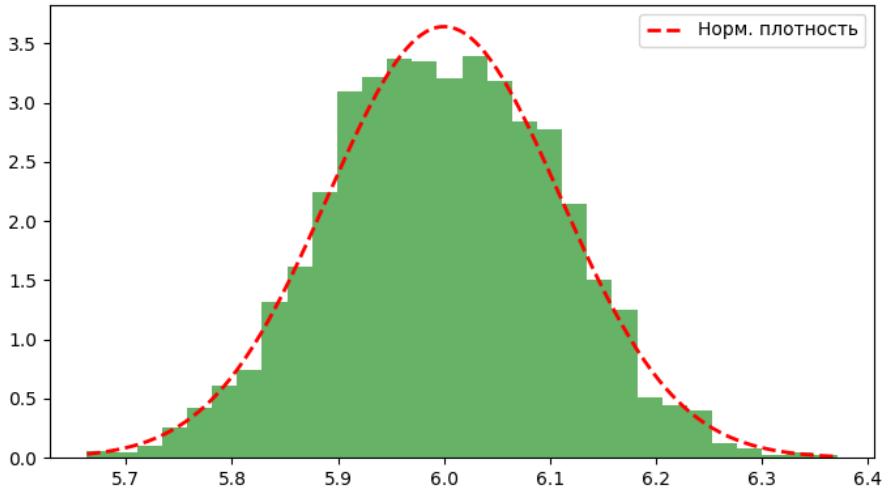

Рис. 1: Гистограмма выборочного среднего.

Как видно на рис. 1, гистограмма имеет выраженную колоколообразную форму и хорошо согласуется с наложенной нормальной кривой. Пик распределения расположен близко к теоретическому среднему, а «хвосты» симметричны относительно этого пика.

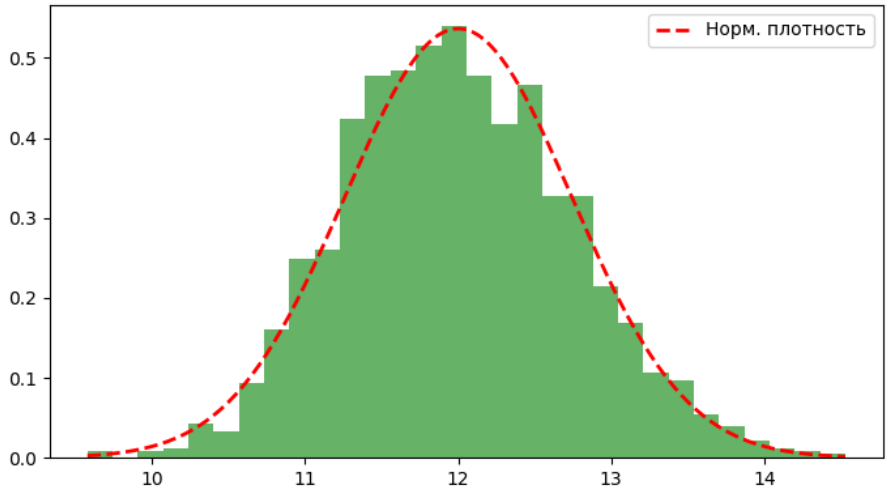

Рис. 2: Гистограмма выборочной дисперсии.

Распределение выборочной дисперсии тоже выглядит близким к нормальному, хотя иногда оно бывает чуть более «асимметричным», чем выборочное среднее. Максимум гистограммы находится около теоретической дисперсии, и «хвосты» распределения сравнительно симметричны.

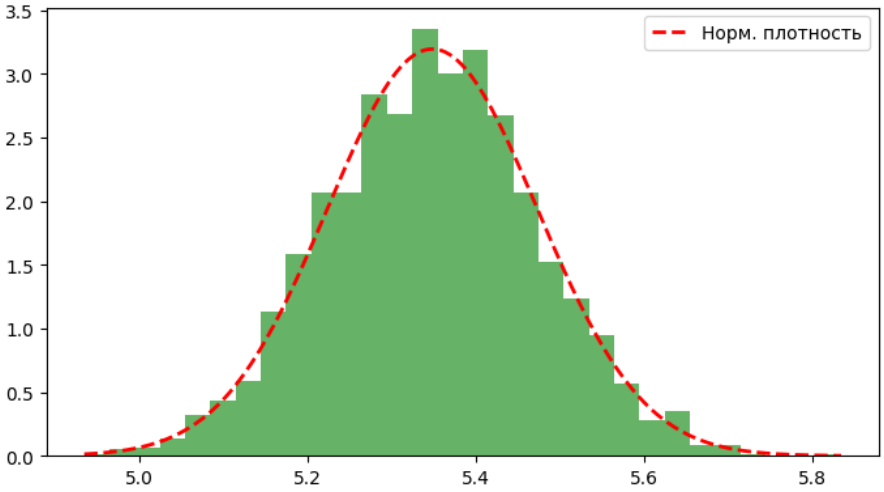

Рис. 3: Гистограмма выборочной квантили порядка 0,5.

Гистограмма медианы показывает, что центр графика находится рядом с истинной медианой, что означает, что при большом объёме выборки медиана почти всегда получается около этого значения.

Вычислим теоретические значения мат. ожидания, дисперсии и медианы для гамма-распределения $\Gamma(3, 2)$ с параметрами $\alpha = 3$ и $\theta = 2$. Тогда:

$$\mathbb{E}[X] \;=\; \alpha\,\theta \;=\; 3 \times 2 \;=\; 6,$$

$$\mathrm{D}[X] \;=\; \alpha\,\theta^2 \;=\; 3 \times (2)^2 \;=\; 12.$$

Точная формула для медианы гамма-распределения в элементарных функциях не выражается, поэтому её находят через обратную функцию распределения (квантильную функцию) при уровне 0.5:

$$\mathrm{median}[X] \;=\; F^{-1}(0.5).$$

С использованием пакета `scipy.stats` в Python вычисляем приближенное значение:

$$\mathrm{median}[X] \;\approx\; 5.348.$$

Выведем численные результаты мат. ожидания, дисперсии и медианы для анализа:

	$\overline{X}$	S^2	median
Среднее по M	5.999	11.991	5.348
Дисперсия по M	0.012	0.573	0.016
Медиана по M	5.999	11.957	5.348
Теор. значение	6	12	5.348

Таблица 1: Числовые результаты по распределениям оценок

Все три гистограммы (для выборочного среднего, выборочной дисперсии и медианы) демонстрируют «колоколообразную» форму, хорошо согласующуюся с наложенной нормальной плотностью. Это свидетельствует об асимптотической нормальности рассмотренных оценок. Дополнительно, таблица с численными результатами показывает, что средние

значения этих оценок (по всем сгенерированным выборкам) близки к теоретическим (мат. ожидание около 6, дисперсия около 12, медиана около 5.35). При этом «дисперсия по M» для выборочной дисперсии (около 0.57) отражает разброс самой оценки S^2 между повторными экспериментами, а не расхождение с теоретическим $\mathrm{D}[X] = 12$. Таким образом, и визуальный, и численный анализ подтверждают корректность генерации выборок, согласие с теорией и асимптотическую нормальность статистик.

Теперь необходимо экспериментально убедиться в сходимости следующих величин к гамма-распределениям:

$$U_1 = n\,F\big(X_{(2)}\big) \;\to\; \Gamma(2,1),$$
$$U_2 = n\,\big(1 - F\big(X_{(n)}\big)\big) \;\to\; \Gamma(1,1),$$

где $X_{(2)}$ — вторая порядковая статистика (второе по величине значение в выборке), $X_{(n)}$ — максимум (наибольшее значение выборки), а $F(x)$ — функция гамма-распределения. Для каждой выборки найдем $X_{(2)}$ и $X_{(n)}$, то есть отсортируем выборку и возьмем второй и максимальный элементы. Подставим их в формулы и построим гистограммы для массивов $\{U_1^{(i)}\}$ и $\{U_2^{(i)}\}$. Для визуального сравнения наложим теоретические плотности $\Gamma(2,1)$ и $\Gamma(1,1)$ соответственно.

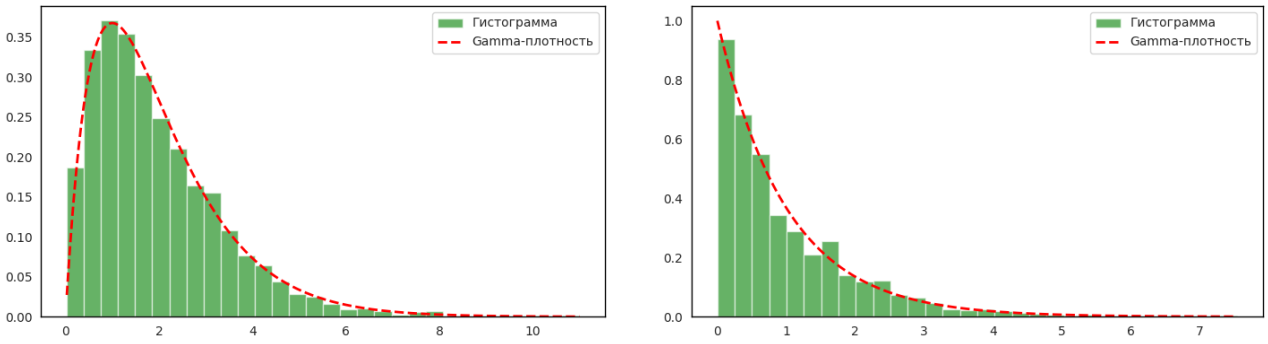

Рис. 4: Гистограммы $U_1 = n\,F(X_{(2)})$ (слева) и $U_2 = n\,(1 - F(X_{(n)}))$ (справа) с наложенными плотностями $\Gamma(2,1)$ и $\Gamma(1,1)$ соответственно.

Как видно из рисунка, в обоих случаях распределения гистограмм близки к теоретическим кривым гамма-распределения. При достаточно большом объёме выборки n и количестве повторений M наблюдается хорошее совпадение с соответствующими плотностями.

Задание 2

В файле `cars93.csv` содержатся данные о 93 моделях автомобилей: их мощность (`Horsepower`), тип (`Type`), производитель (`Manufacturer`) и другие характеристики. Для выполнения задания необходимо загрузить датасет и ответить на следующие вопросы:

1. Определим, какие типы автомобилей (`Type`) представлены в датасете, а также найдем наиболее и наименее распространённый тип.

2. Для мощности (`Horsepower`) вычислить (для всех автомобилей и для каждого типа отдельно):

   - выборочное среднее,
   - выборочную дисперсию,

- выборочную медиану,
- межквартильный размах.

3. Построить (для всех автомобилей и для каждого типа отдельно):
   - график эмпирической функции распределения,
   - гистограмму,
   - box-plot.

С помощью библиотеки `pandas` загрузим датасет и выведем какие типы автомобилей представлены и их количество. Далее найдем наиболее и наименее распространенный тип автомобилей:

Тип	**Количество**
`Midsize`	22
`Small`	21
`Compact`	16
`Sporty`	14
`Large`	11
`Van`	9

Таблица 2: Результаты программы

Как видно по таблице наиболее распространённым типом является Midsize, а наименее распространённый — Van.

Теперь вычислим для мощности (`Horsepower`) выборочное среднее, выборочную дисперсию, выборочную медиану и межквартильный размах для всех автомобилей:

Статистика	**Значение**
$\overline{X}$	143.83
S^2	2743.08
$Median$	140.00
IQR	67.00

Таблица 3: Статистики для всей совокупности

И для каждого типа отдельно:

	`Compact`	`Large`	`Midsize`	`Small`	`Sporty`	`Van`
$\overline{X}$	131.00	179.45	173.09	91.00	160.14	149.44
S^2	518.53	477.07	2756.09	447.60	5536.29	370.28
$Median$	132.00	170.00	169.00	90.00	147.50	151.00
IQR	33.25	25.00	69.00	21.00	72.25	23.00

Таблица 4: Статистики для каждого типа отдельно

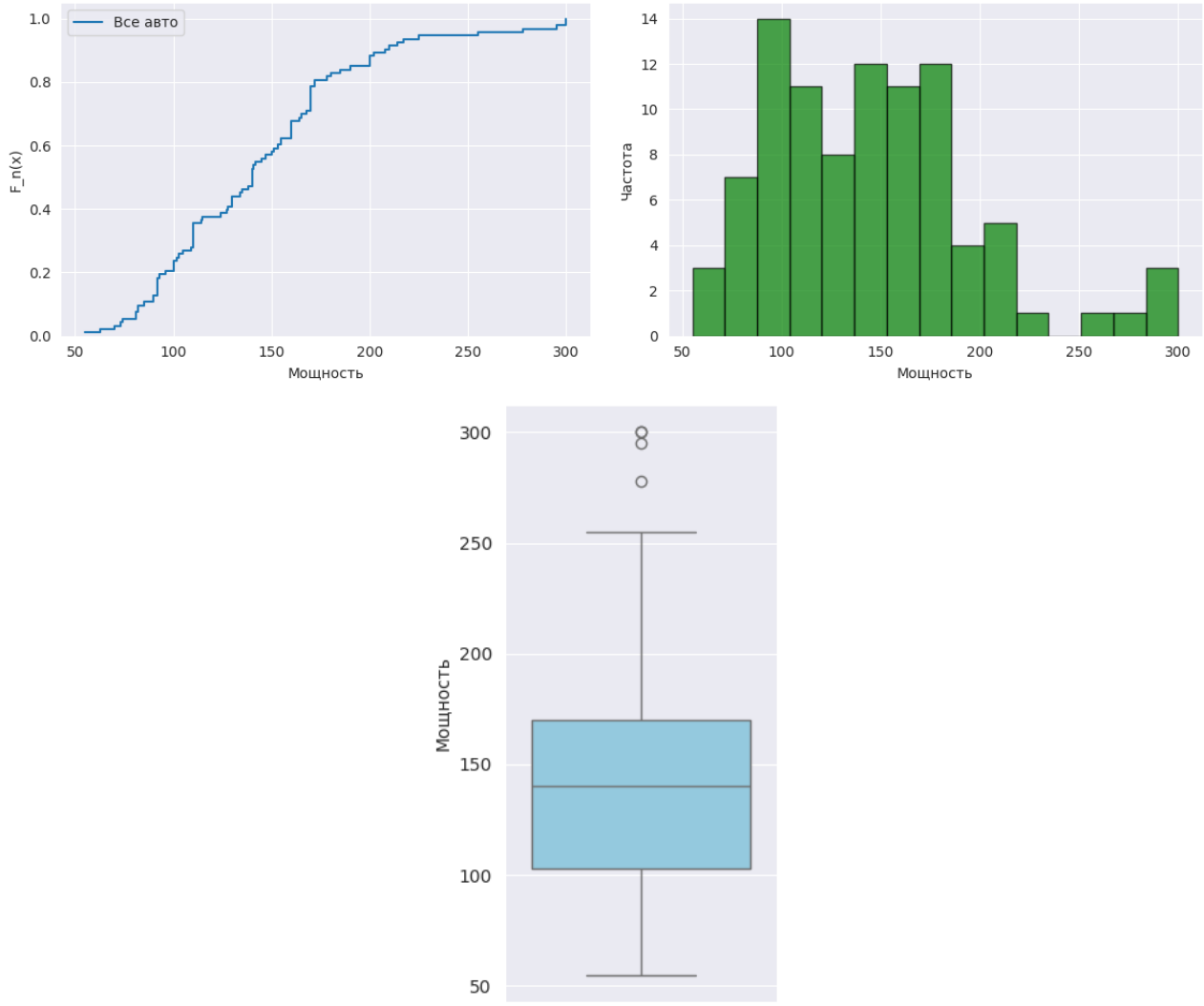

Рис. 5: Графики мощности (`Horsepower`) для всех автомобилей: эмпирическая функция распределения, гистограмма, box-plot.

Построенные графики показывают одну информацию, но в разных визуальных формах. Разберем каждый из них:

- **Эмпирическая функция распределения.** Ступенчатая кривая показывает, как распределены автомобили по мощности. Например, по оси x можно примерно оценить, что около 50 % автомобилей (то есть медиана) имеют мощность порядка 130–140. Видно, что кривая накапливается постепенно до значений около 200 л. с., затем есть относительно небольшой «скачок» к 250–300 л. с., где находятся немногие машины.

- **Гистограмма.** Наиболее массовый диапазон мощностей — примерно 90–170 л. с. Есть «пик» в районе 100–110 л. с. и 140–150 л. с., а также заметный разрыв в интервале 210–240 л. с. Небольшое количество автомобилей имеет мощность выше 250 л. с.

- **Box-plot.** Отчётливо видно, что медиана находится примерно в диапазоне 130–140 л. с. Верхняя и нижняя границы «ящика» соответствуют первому и третьему квартилю. В данной выборке есть несколько выбросов (точки сверху), то есть модели с существенно большей мощностью (более 250 л. с.).

Теперь построим графики распределения мощности по типам автомобилей

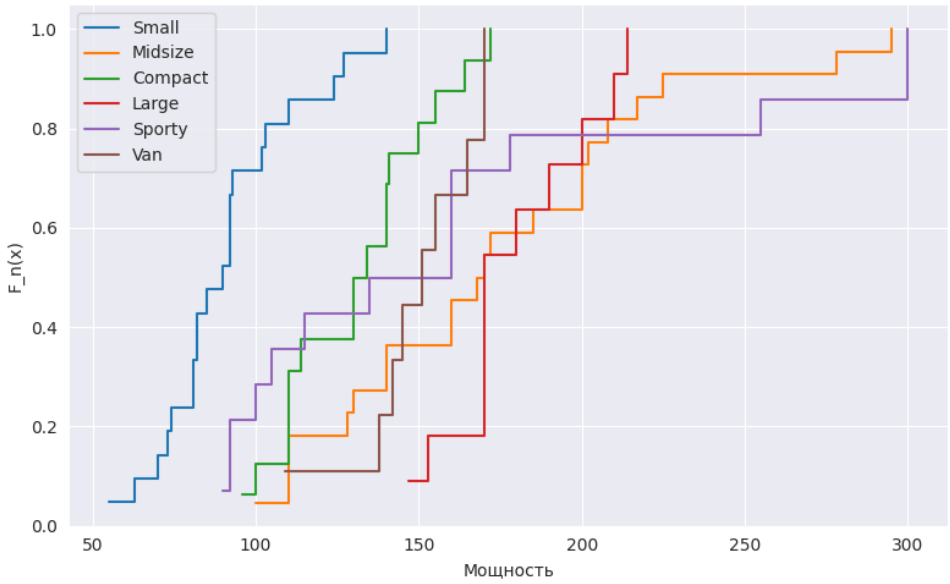

Рис. 6: Эмпирические функции распределения мощности по каждому типу авто.

На рис. 6 представлены эмпирические функции распределения для каждого типа:

- Кривая для `Small` смещена влево (большая часть значений ниже 120 л. с.).
- `Midsize` и `Sporty` имеют более «длинный хвост» вправо: некоторые модели достигают 250–300 л. с.
- `Compact` и `Van` занимают промежуточное положение между `Small` и `Large`.

Ниже приведены гистограммы для каждого типа автомобиля по отдельности, позволяющие более детально увидеть характер распределения внутри группы.

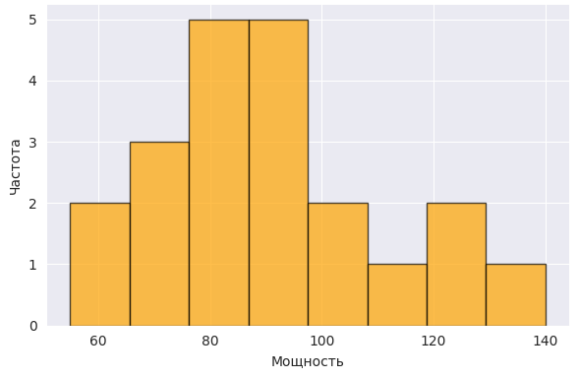

Рис. 7: `Small`

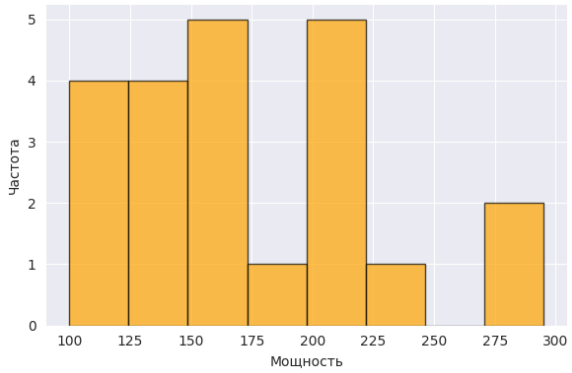

Рис. 8: `Midsize`

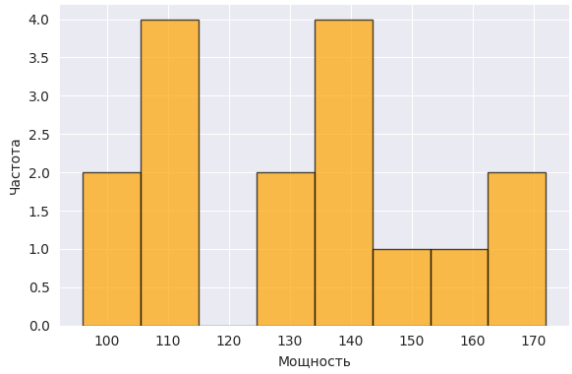

Рис. 9: `Compact`

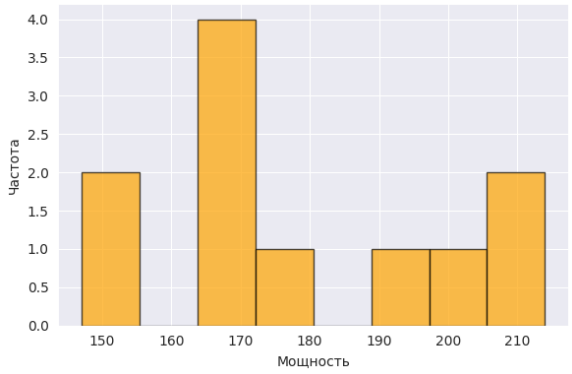

Рис. 10: `Large`

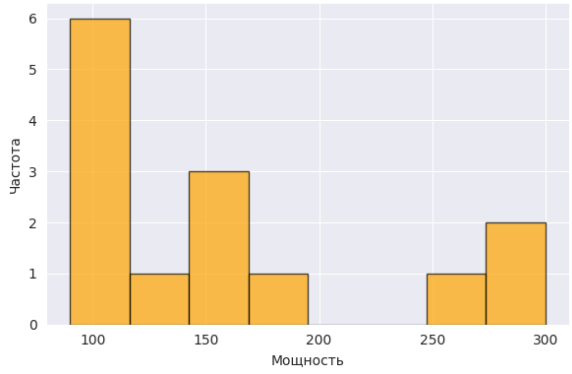

Рис. 11: `Sporty`

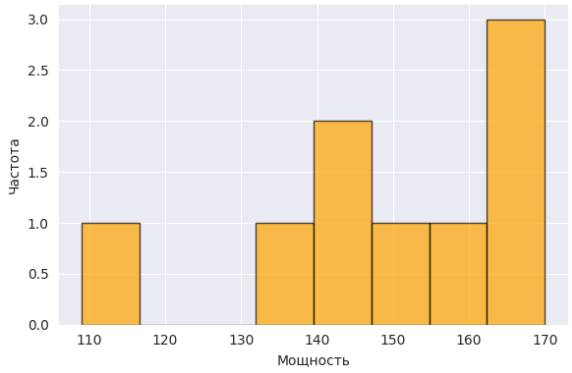

Рис. 12: `Van`

Гистограммы показывают, что распределение мощности внутри каждого типа автомобиля имеет свои особенности:

- Типы `Small` и `Compact` характеризуются сравнительно невысокой мощностью (часто менее 140–150 л. с.), причём у `Small` заметен минимум около 50 л. с.
- `Midsize`, `Large` и `Sporty` имеют более широкий разброс: есть модели как с мощностью около 100 л. с., так и с показателями свыше 200–250 л. с.
- У `Van` мощность в основном сосредоточена в интервале 130–170 л. с. и редко выходит за его пределы.
- В целом, если сравнить медианы (см. box-plot), наиболее «слабомощными» оказываются автомобили `Small`, а наиболее «сильными» — отдельные модели из `Midsize` и `Sporty`.

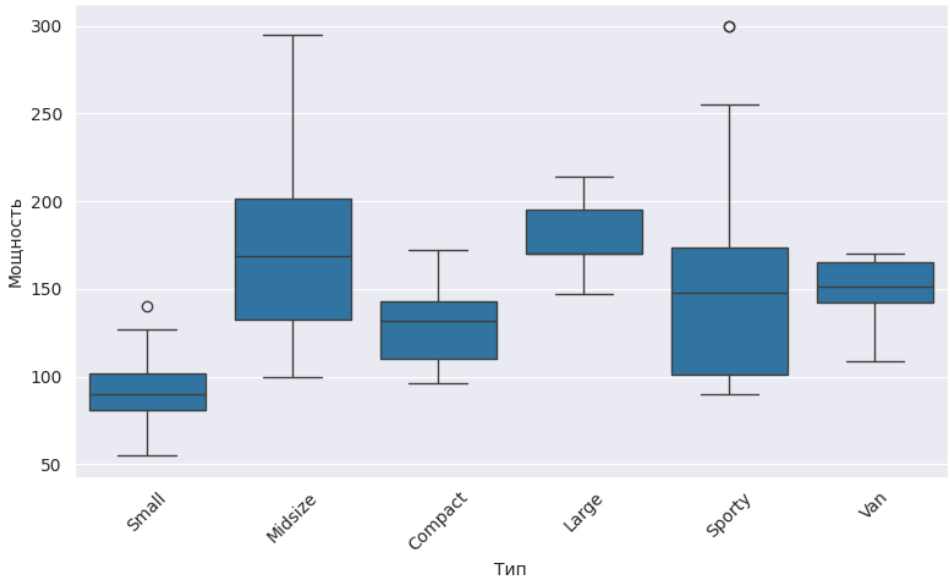

Рис. 13: Box-plot распределения мощности по каждому типу автомобиля.

На рис. 13 видно, что распределение мощности существенно варьируется в зависимости от типа авто. Например:

- `Small` — имеют сравнительно невысокую мощность (медиана около 90–100 л. с.), при этом присутствует один автомобиль с ещё более низкой мощностью (выброс).
- `Midsize` — разброс достаточно велик, верхняя «уса» достигает 300 л. с., а медиана лежит в районе 170 л. с.
- `Compact` — в среднем чуть меньше 140 л. с., заметно более «скромный» разброс, чем у `Midsize` или `Sporty`.
- `Large` — медиана около 170–180 л. с. и довольно короткие усы, т. е. большая часть значений сгруппирована ближе к центру.
- `Sporty` — имеет значительную вариативность: нижняя граница может быть порядка 90–100 л. с., тогда как верхние значения превышают 250 л. с. (выброс).
- `Van` — мощность в среднем около 150 л. с. с довольно узким разбросом.

Вывод

По результатам выполнения двух заданий можно сделать следующие выводы:

1. **Задание 1.**
   Было выбрано распределение (гамма-распределение с заданными параметрами), у которого существуют первые четыре момента. Сгенерированы большие выборки и экспериментально проверены:
   - Асимптотическая нормальность выборочного среднего, дисперсии и медианы. Построенные гистограммы с наложением нормальных кривых показали хорошее соответствие при достаточно больших размерах выборки.

- Сходимость величин $U_1 = n\,F(X_{(2)})$ к $\Gamma(2,1)$ и $U_2 = n\,(1 - F(X_{(n)}))$ к $\Gamma(1,1)$. Гистограммы этих показателей убедительно совпали с теоретическими плотностями соответствующих гамма-распределений.

Данные результаты подтвердили основные теоретические утверждения о распределениях оценок и порядковых статистик, иллюстрируя, как при увеличении объёма выборки оценки ведут себя согласно предельным теоремам.

2. **Задание 2.**
Использовался датасет `cars93.csv`, содержащий информацию о 93 моделях автомобилей. Для признака `Horsepower` (мощность) были:

   - Определены типы автомобилей (`Type`) и их частотное распределение. Установлено, какой тип встречается чаще всего и какой реже.
   - Вычислены выборочные средние, дисперсии, медианы и межквартильный размах (IQR) мощности как по всей совокупности автомобилей, так и отдельно для каждого типа.
   - Построены гистограммы, эмпирические функции распределения (ECDF) и box-plot, что позволило наглядно сравнить распределения мощности между разными типами авто (например, `Small`, `Midsize`, `Sporty` и т. д.). Показано, что некоторые классы (`Sporty`, `Midsize`) имеют более широкий разброс и могут достигать значений мощности свыше 250–300 л. с., в то время как `Small` и `Compact` концентрируются в более низком диапазоне.

Такой анализ дал представление о структуре реальных данных и подтвердил возможность применять методы описательной статистики и визуализации (гистограммы, box-plot, ECDF) для исследования закономерностей в выборке.