

Математическая статистика

Семинар 1

ФБМФ/ФМХФ МФТИ, 2020
подготовил Андрей Сонин

План занятия

- Что такое математическая статистика
- Зачем она нужна
- Как использовать язык Python 3 в статистике
- Как читать, обрабатывать и хранить данные
- Как отображать информацию о данных

Статистика

Что это такое и зачем она нужна

Описательная

Графическое представление
данных

Расчёт статистических
параметров

Выбор формата хранения
данных

Оценочная

Оценка значений физических
величин

Расчёт достоверности этих
оценок

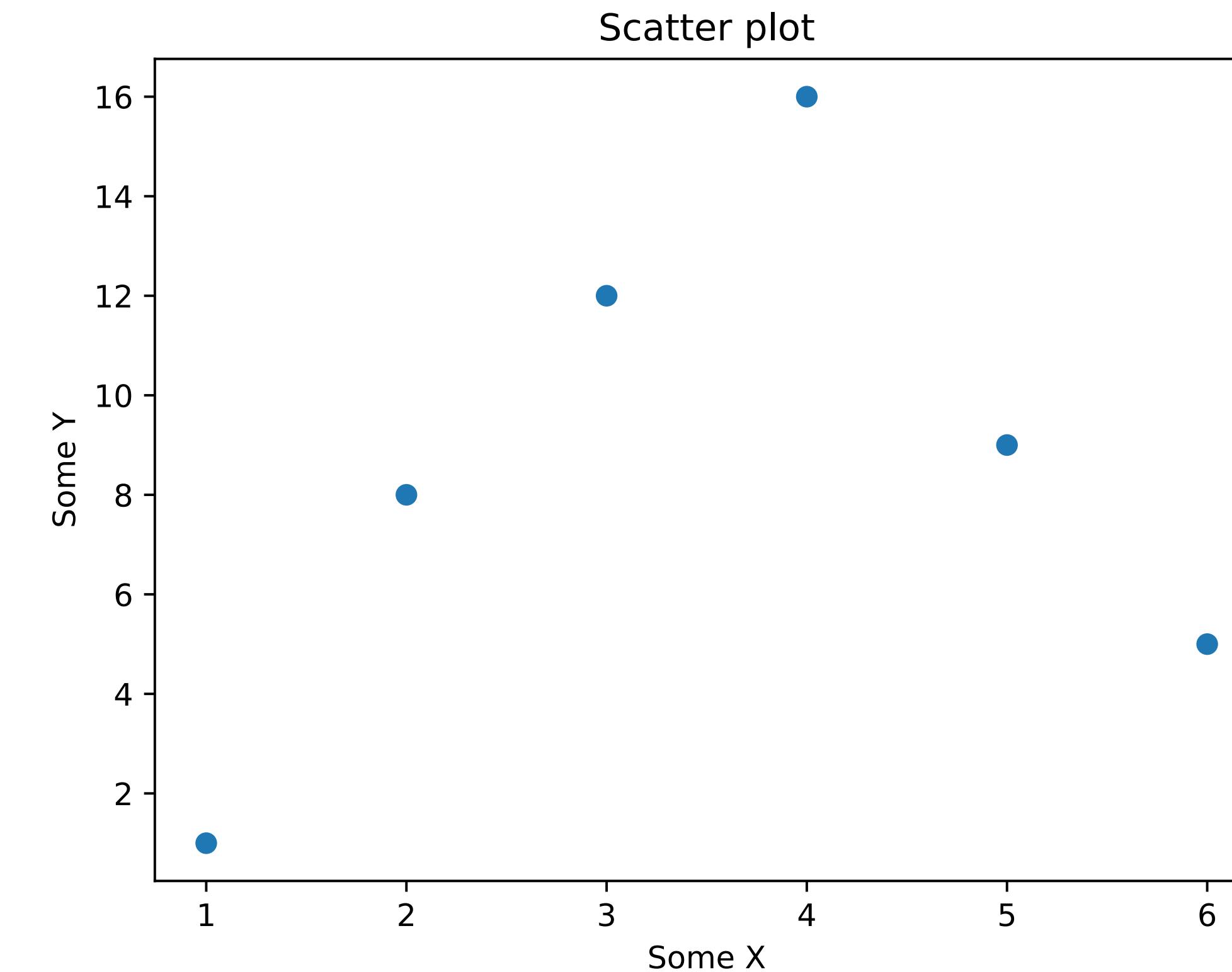
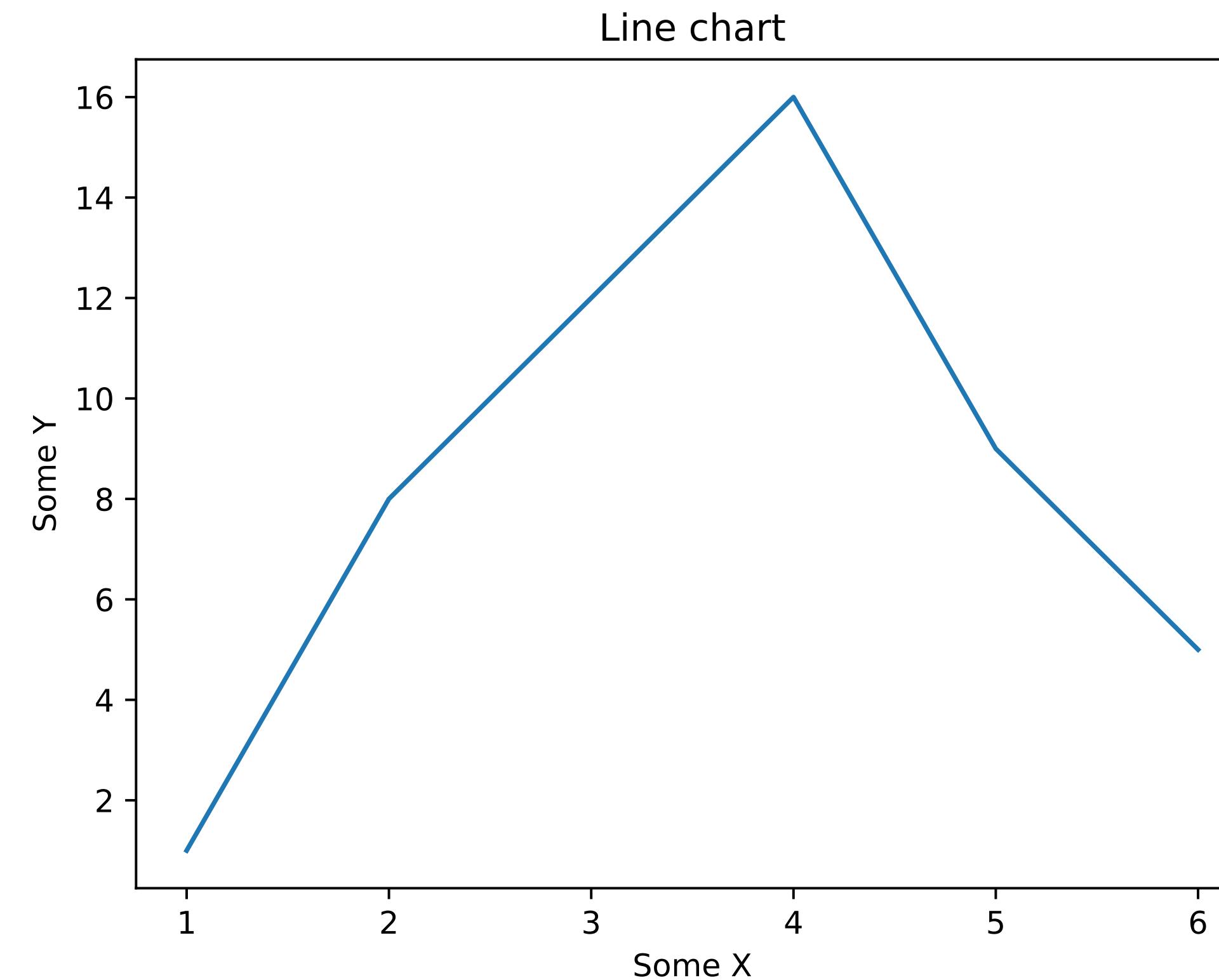
Теория проверки гипотез

Составление гипотез о данных

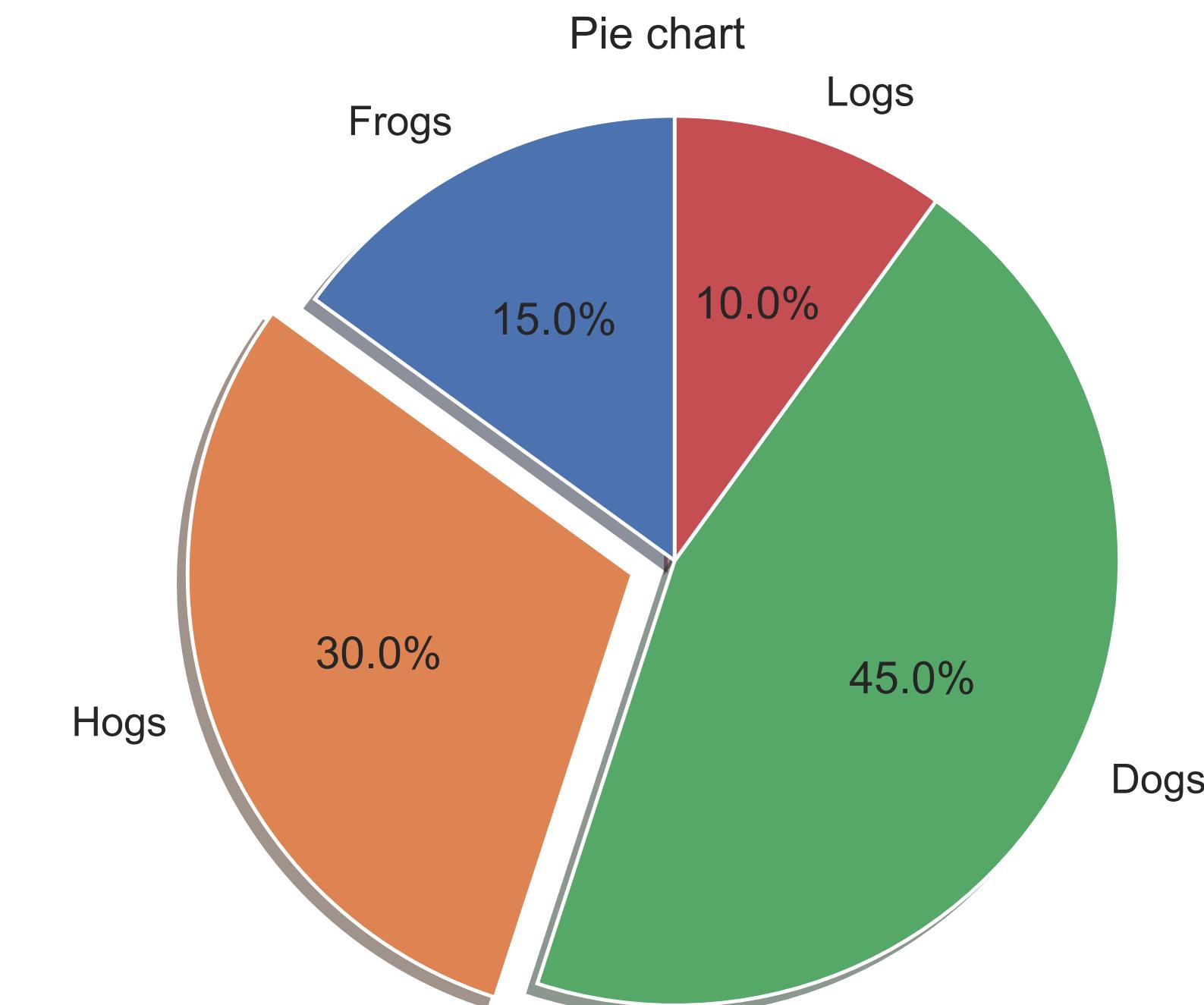
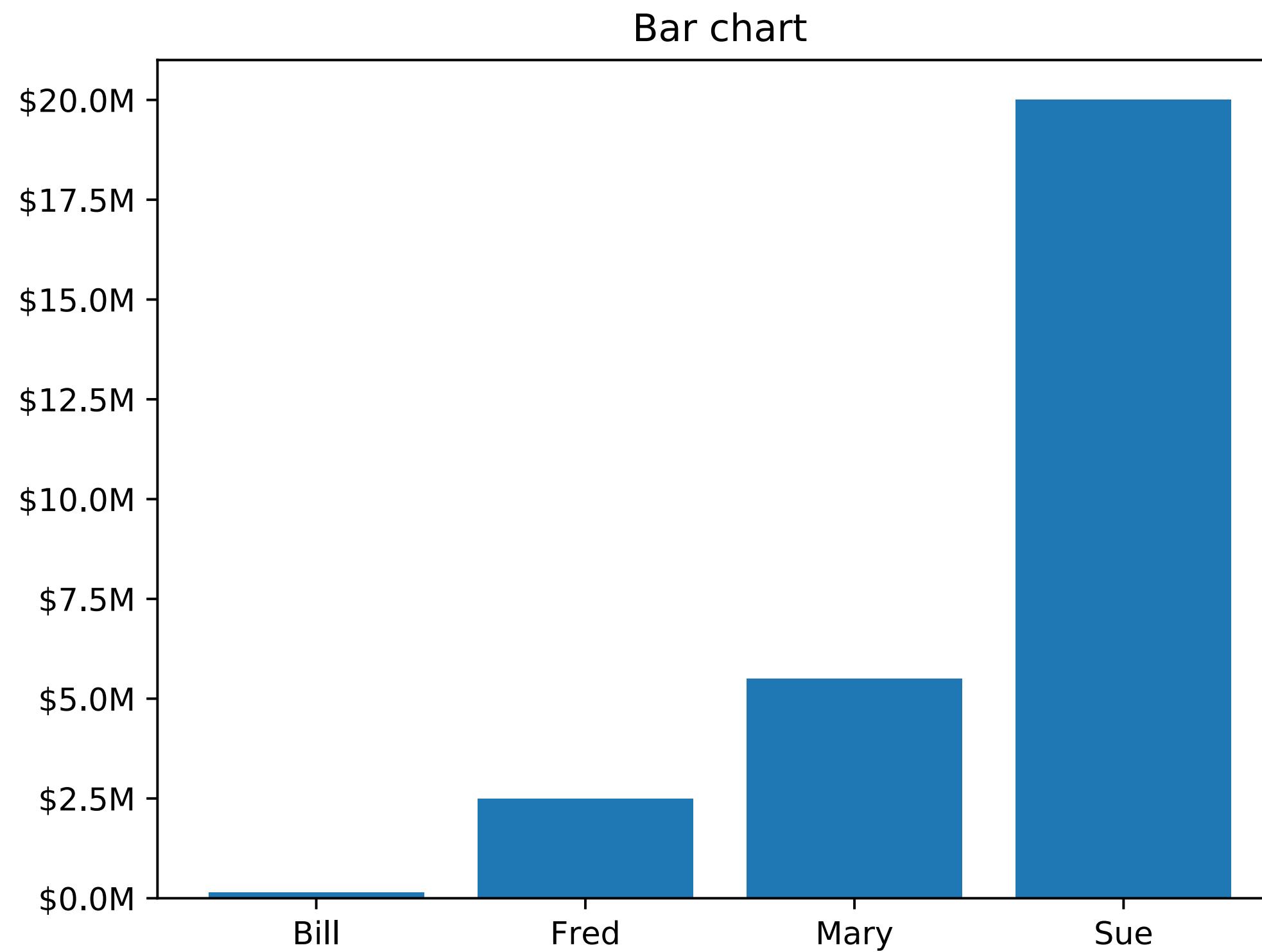
Оценка их справедливости

Расчёт достоверности этих
оценок

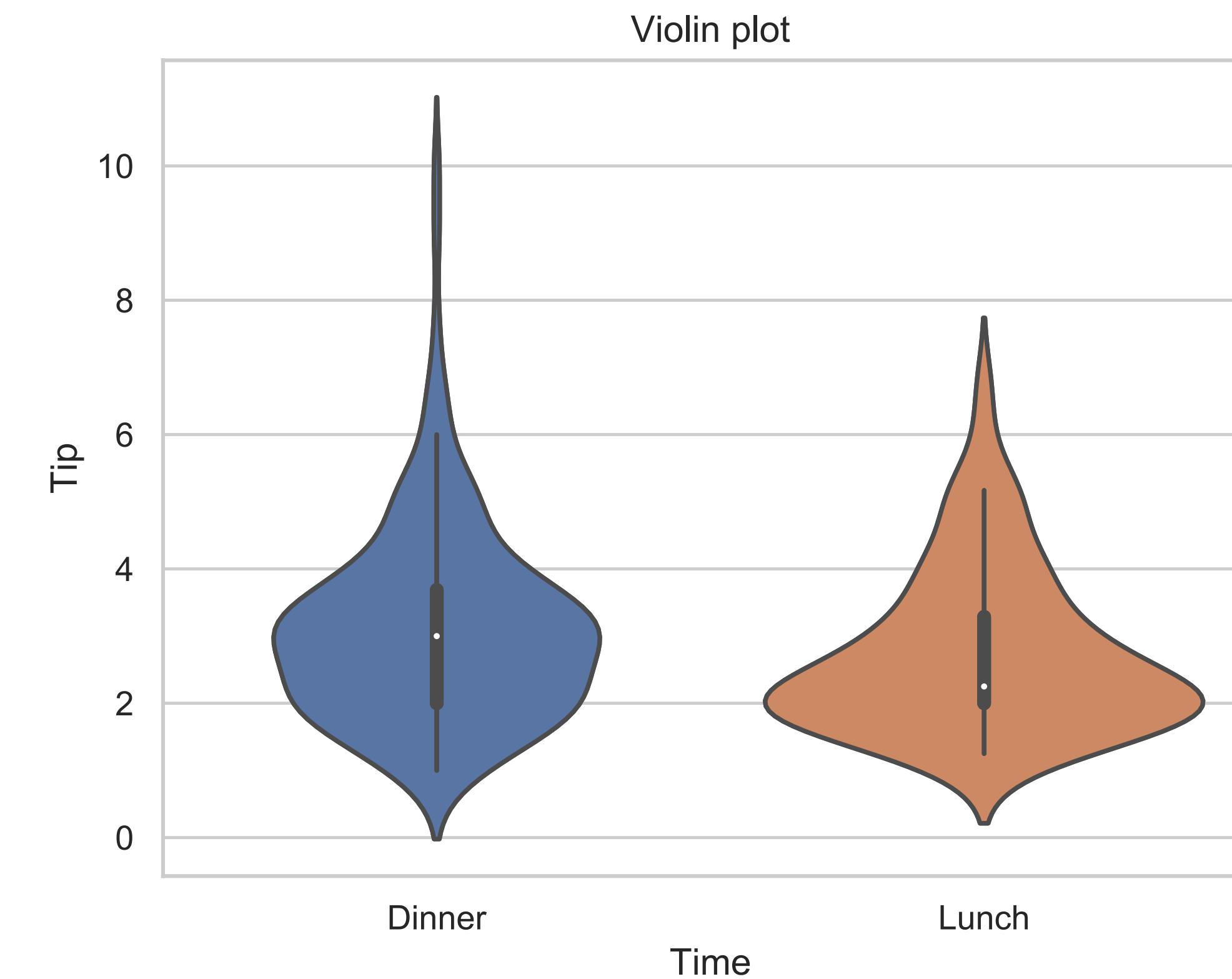
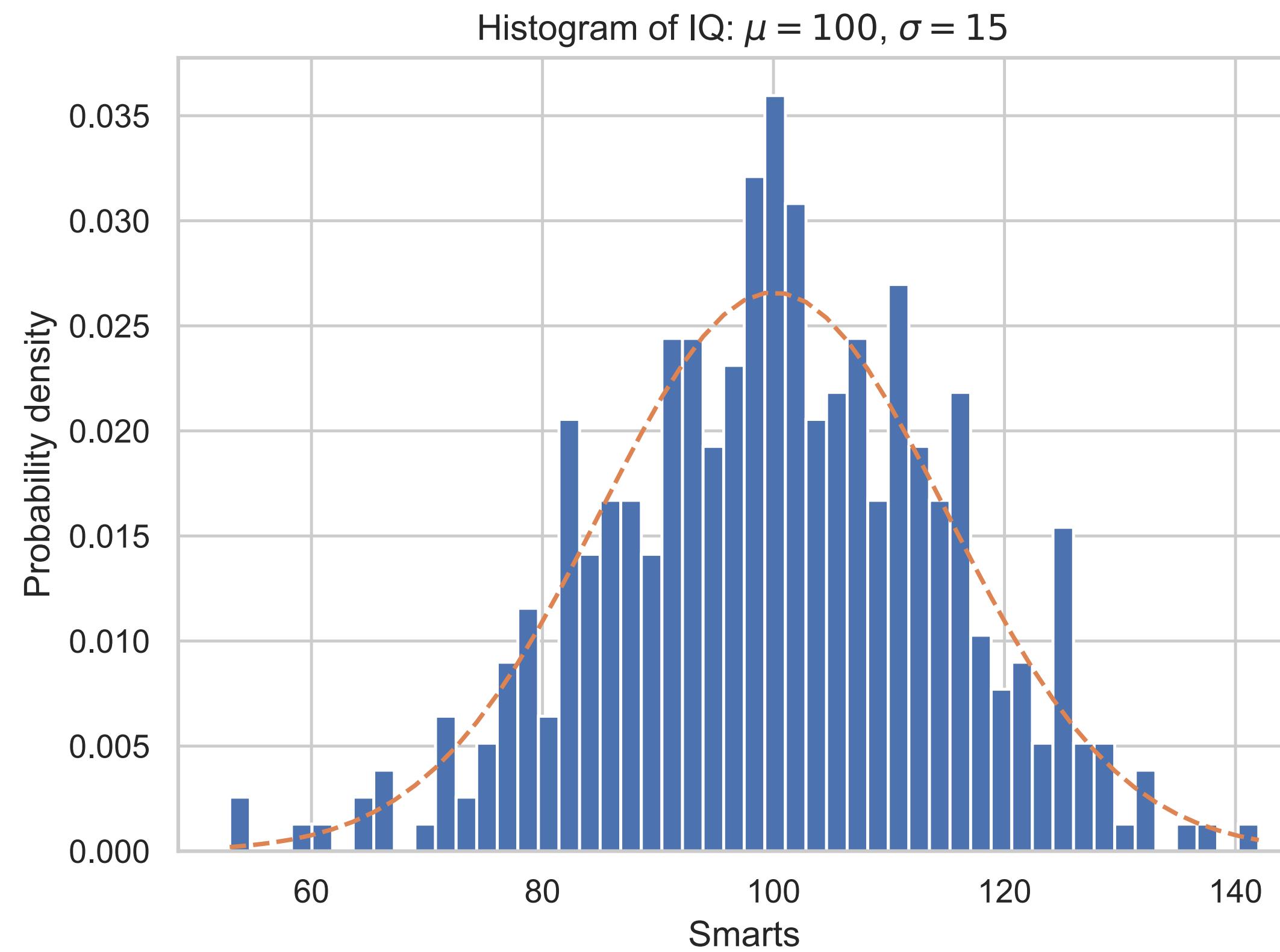
Виды графиков. Простые



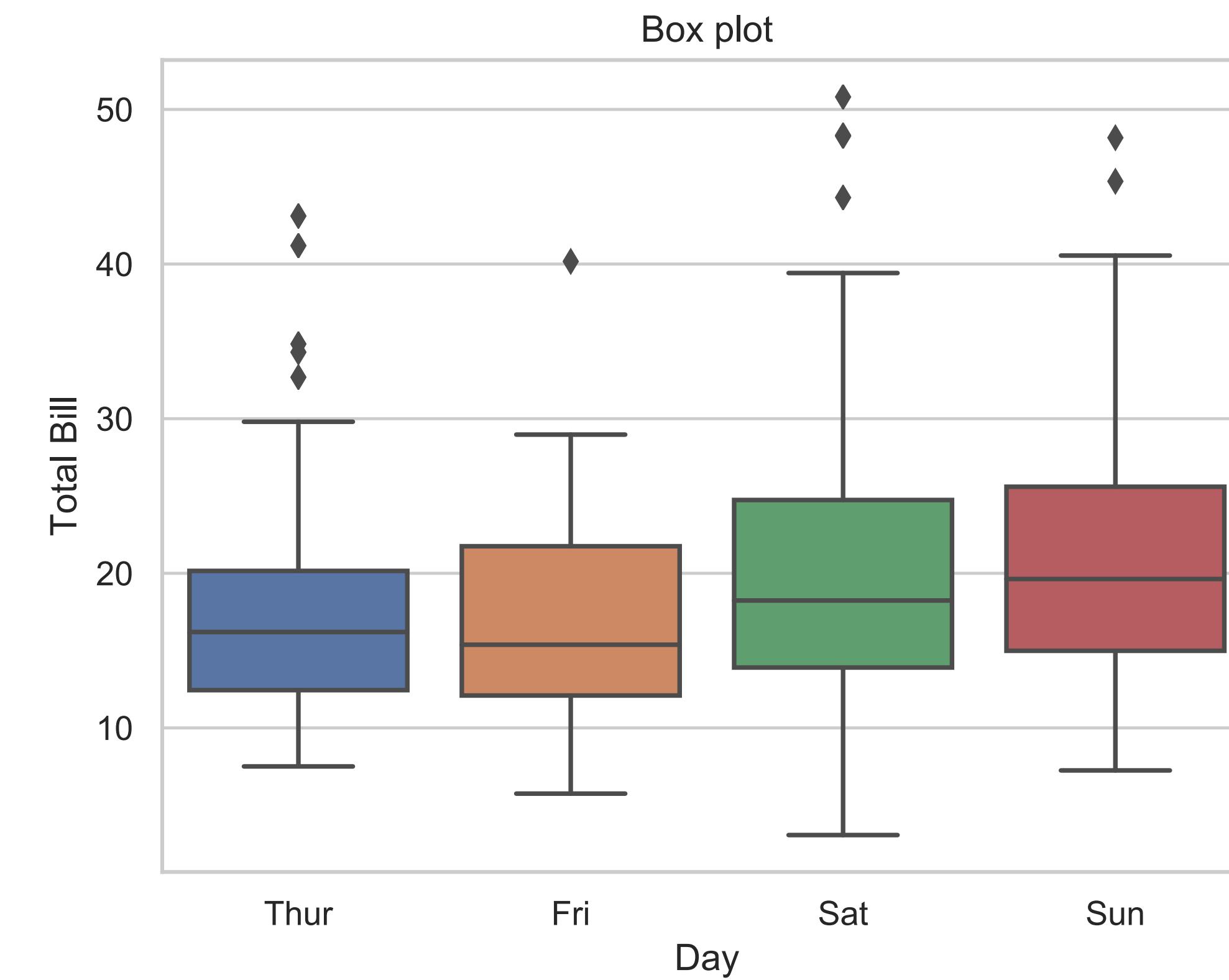
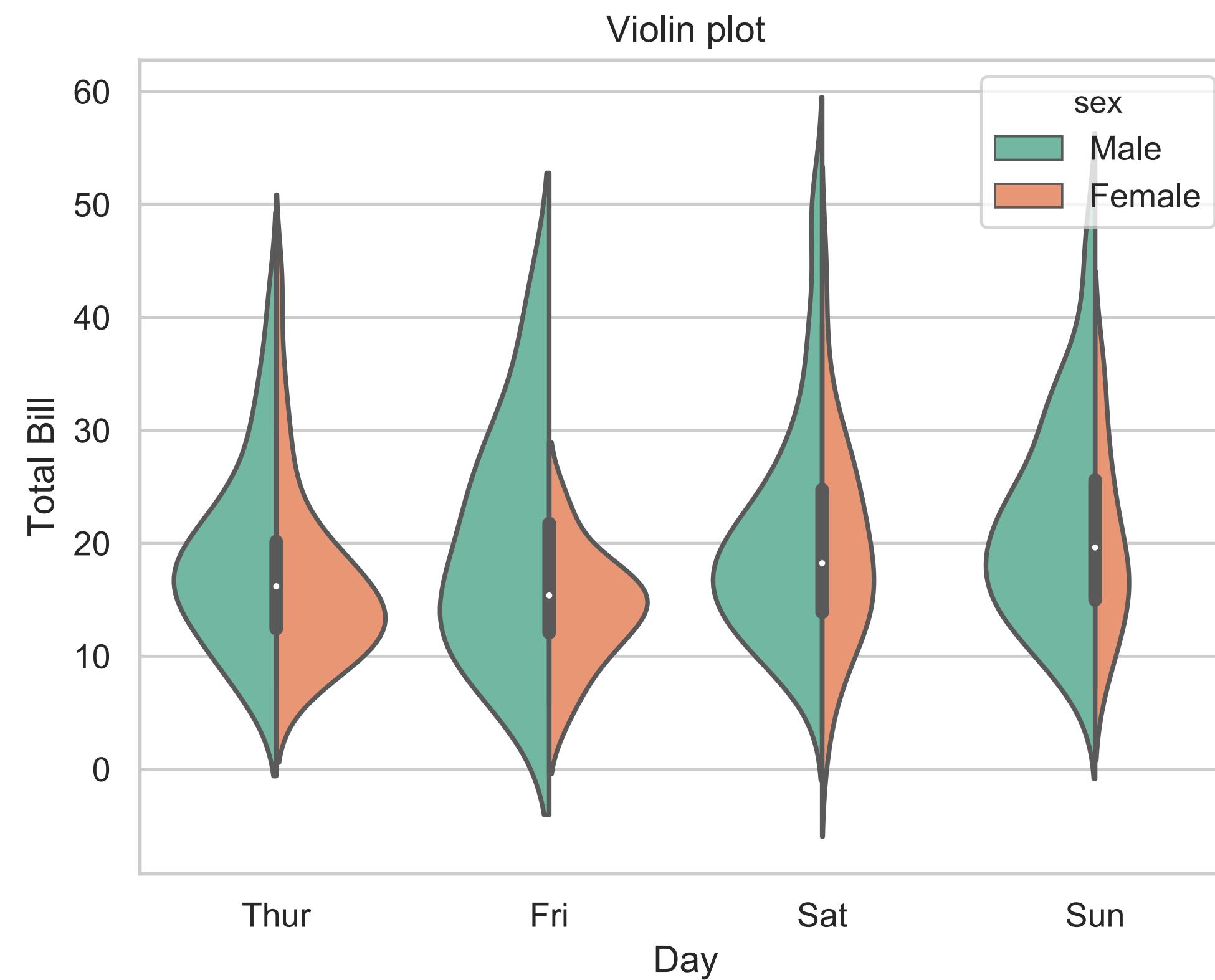
Виды графиков. Категориальные



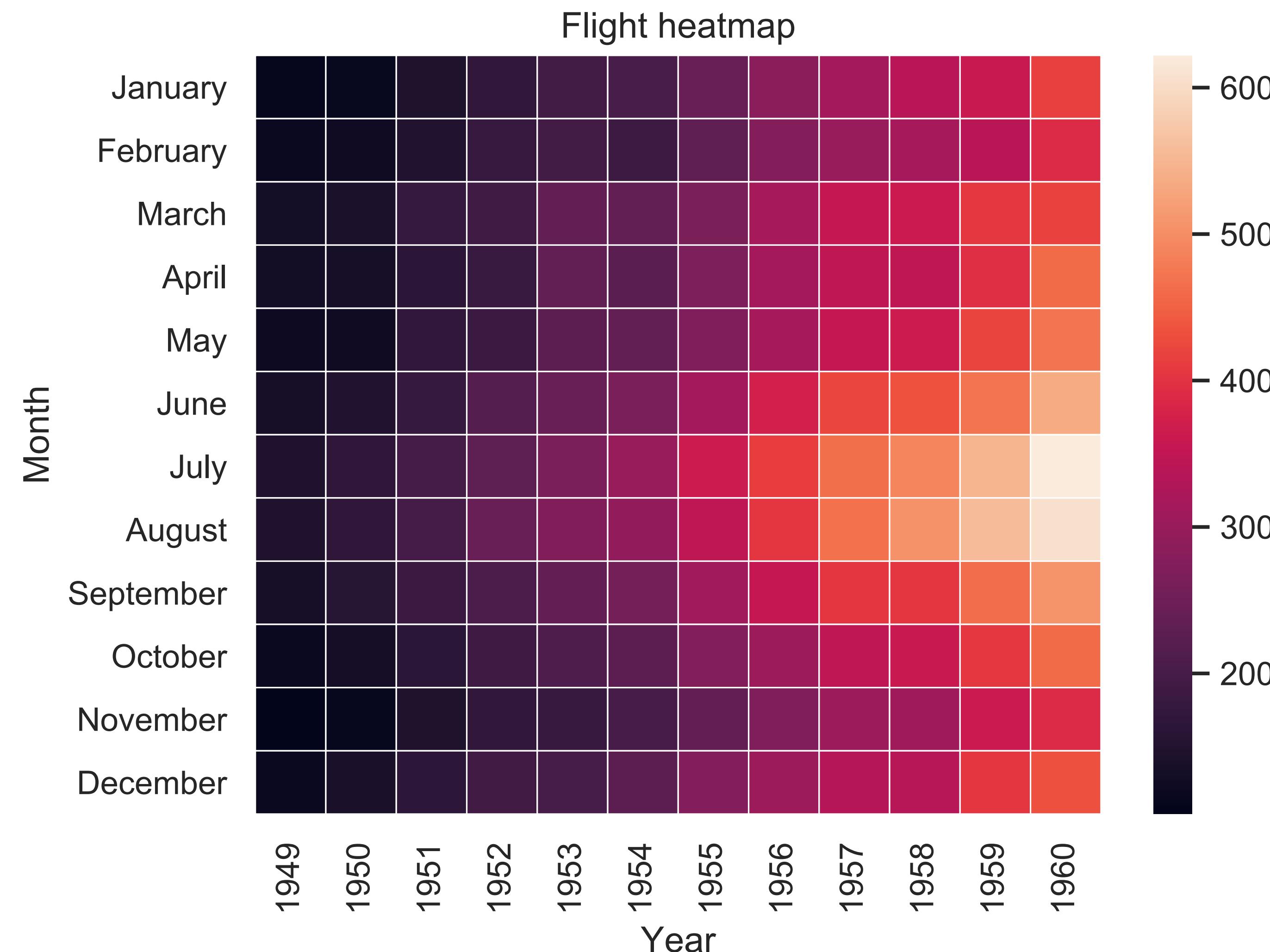
Виды графиков. Распределения



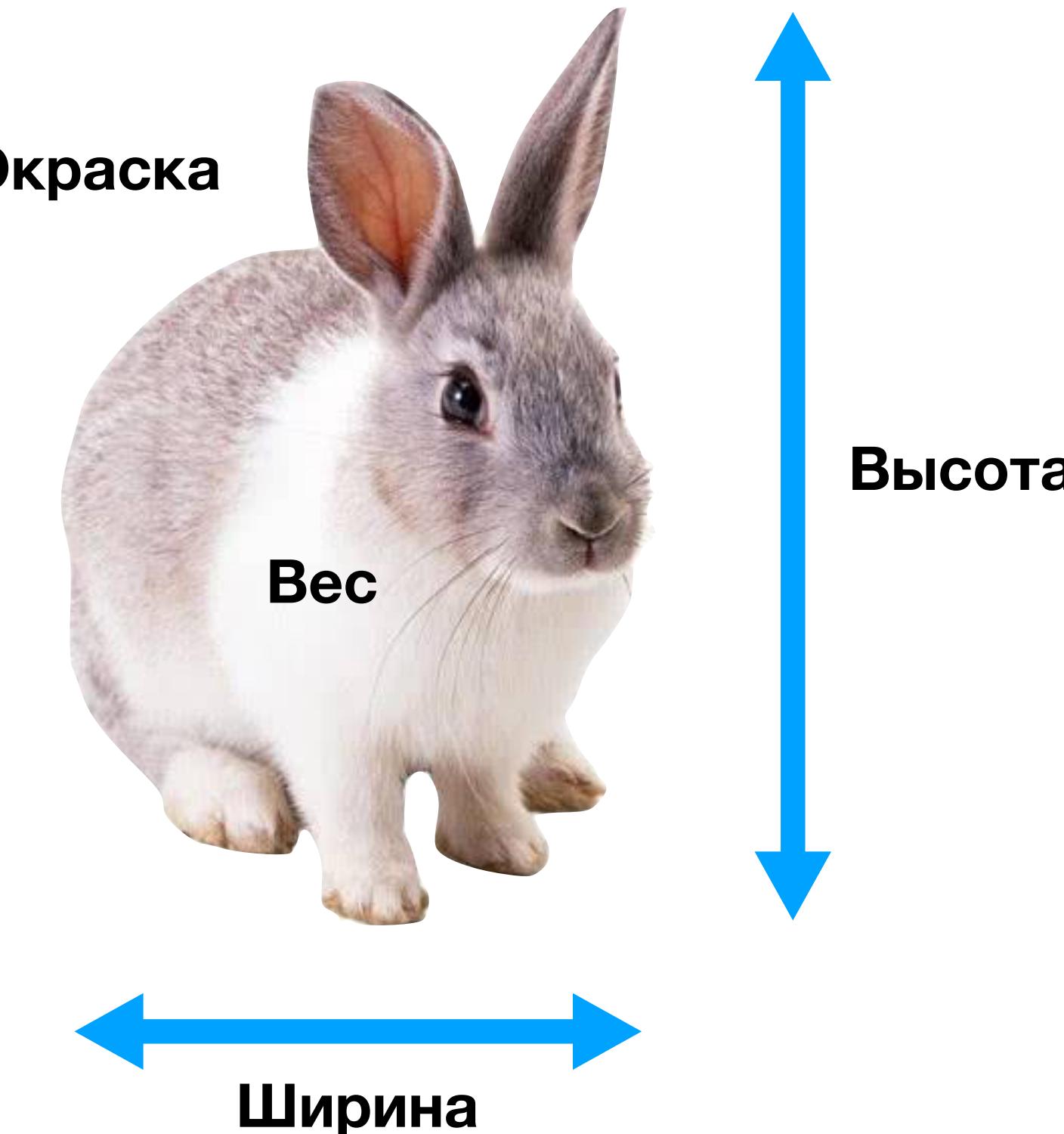
Виды графиков. Распределения



Виды графиков. Тепловая карта



Описательная статистика



Описательная статистика

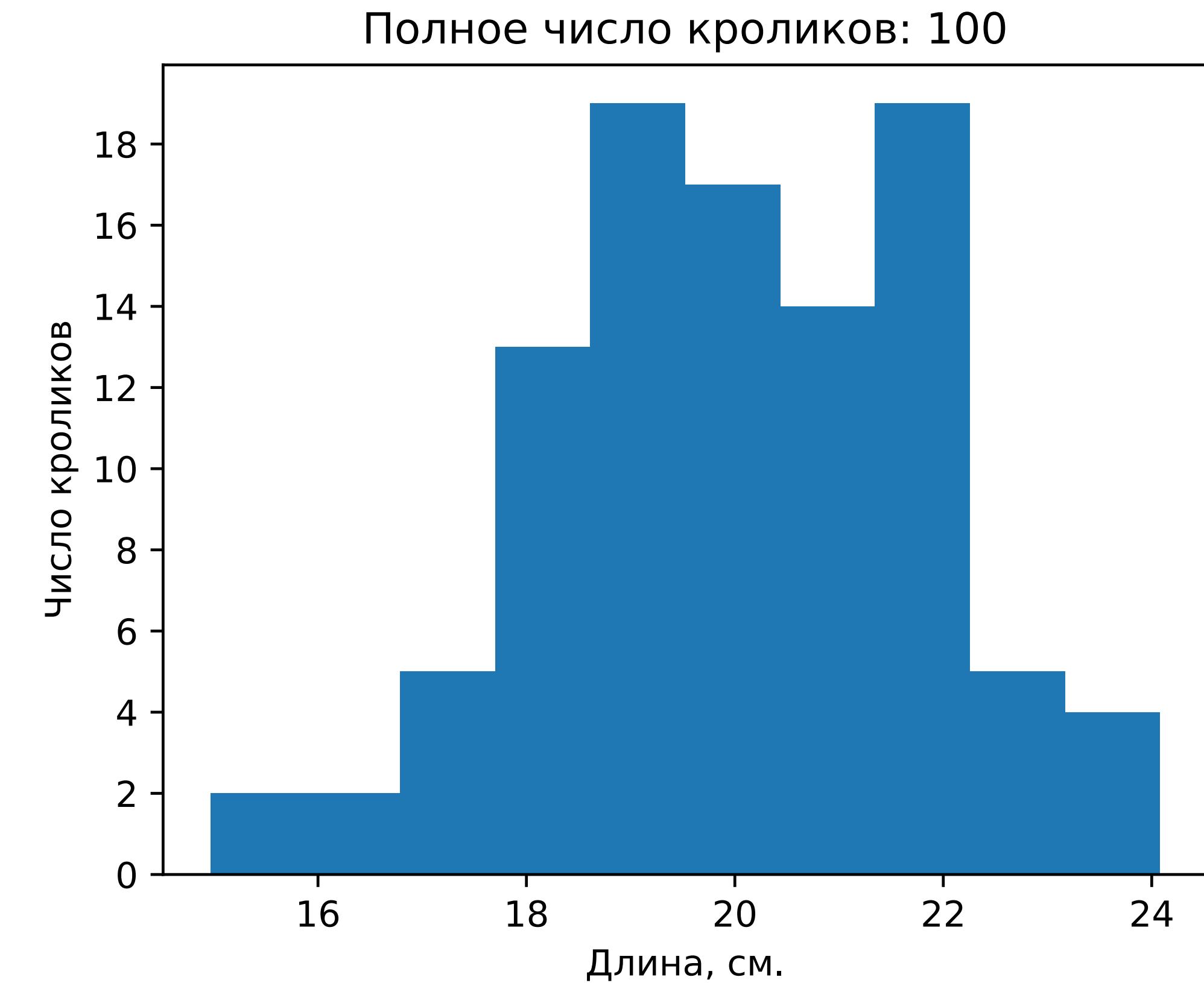


№	Ширина, см.	Высота, см.	Вес, г.	Окраска	Цвет
Кролик 1	20	20	300	Пятнистая	—
Кролик 2	30	40	500	Пятнистая	—
Кролик 3	15	22	450	Пятнистая	—
Кролик 4	22	15	310	Однородн ая	Чёрный

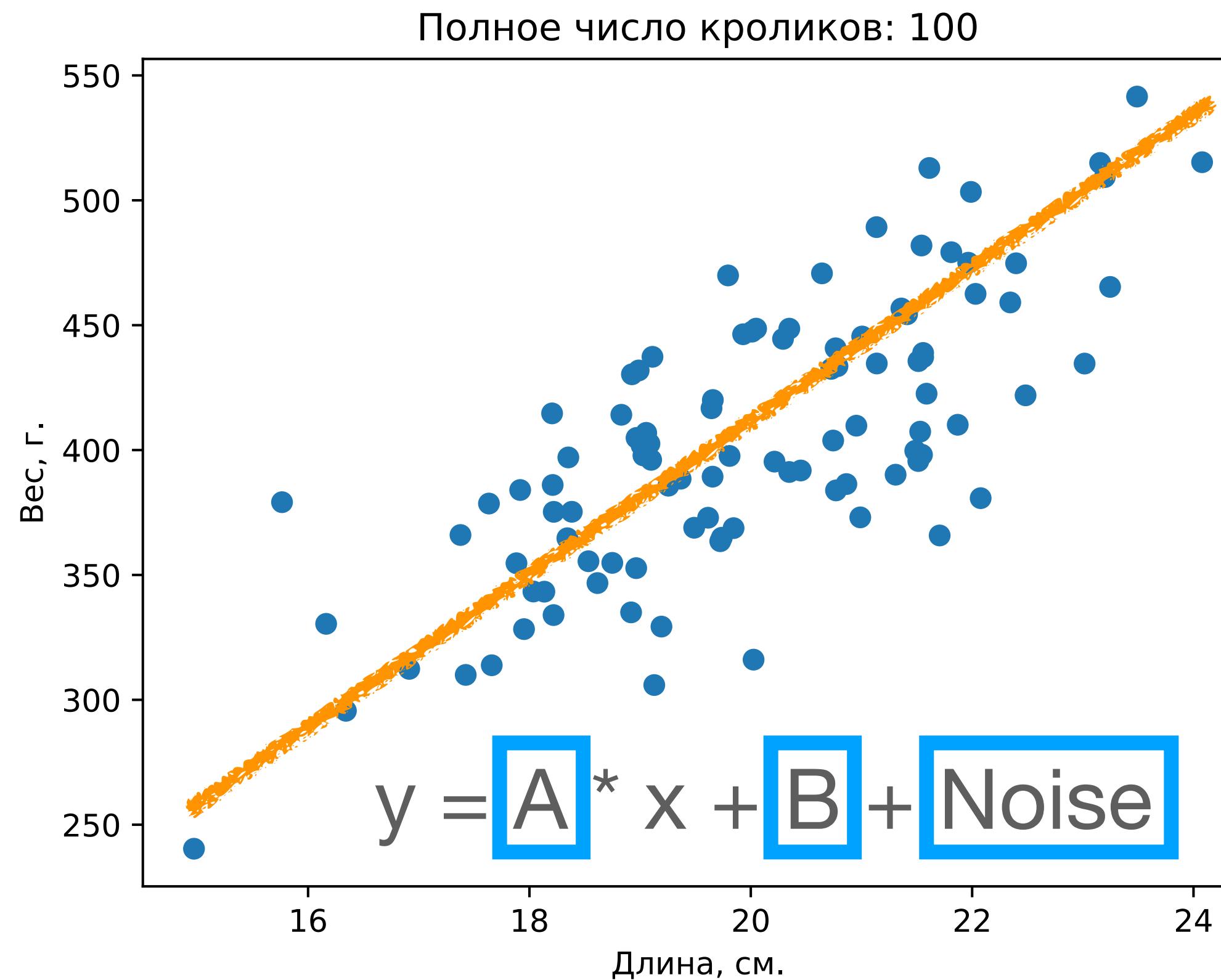
Описательная статистика



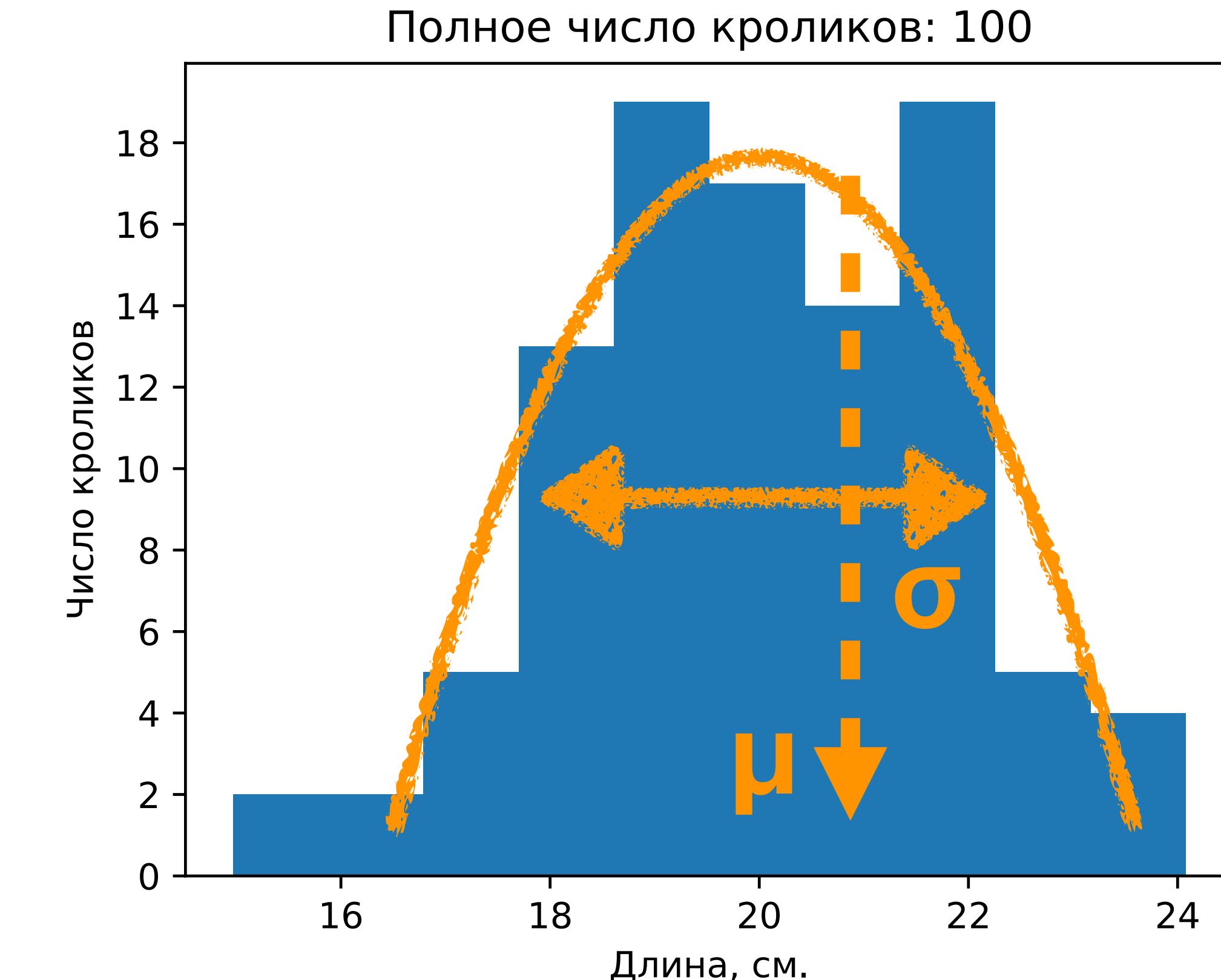
**А что если кроликов
достаточно много?**



Оценочная статистика



$A = 20, B = 0, \text{Noise} \sim \text{Norm}(0, \sigma)$
ДИ для σ : [20, 50]



$\mu = 21 \text{ см}, \sigma = 1.5$
ДИ для σ : [0.8, 2] см

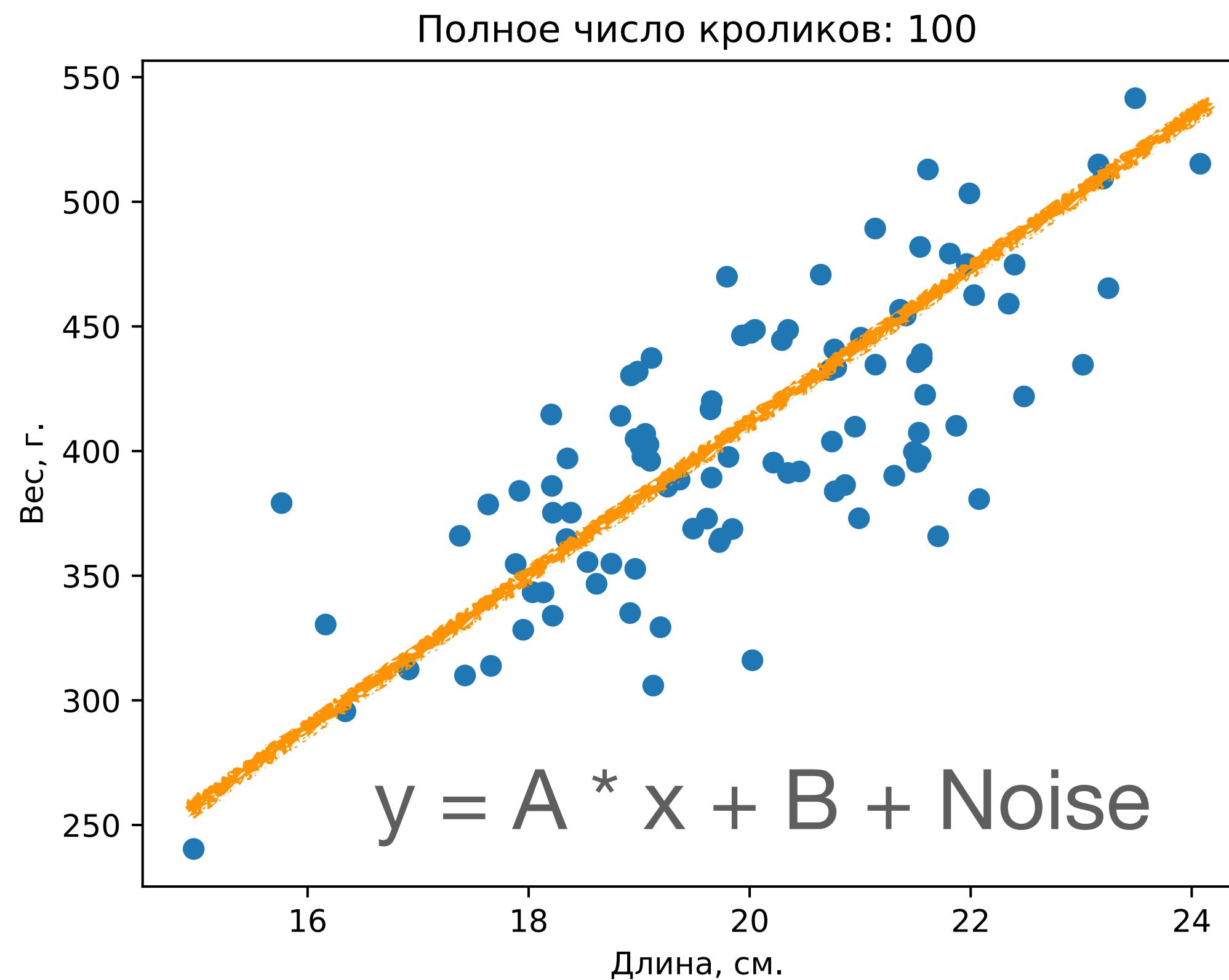
Оценочная статистика

- **Описание** данных не требует априорных предположений о них
- **Оценка** параметров (и их доверительных интервалов) распределений и физических величин уже может требовать постановки гипотез
- Например
 - Эти данные пришли из нормального распределения и нам нужно построить ДИ для дисперсии
 - А эти два признака связаны между собой линейно с шумом определённого вида (и так же требуется оценить ДИ дисперсии шума)
- Но вдруг наши гипотезы ошибочны?

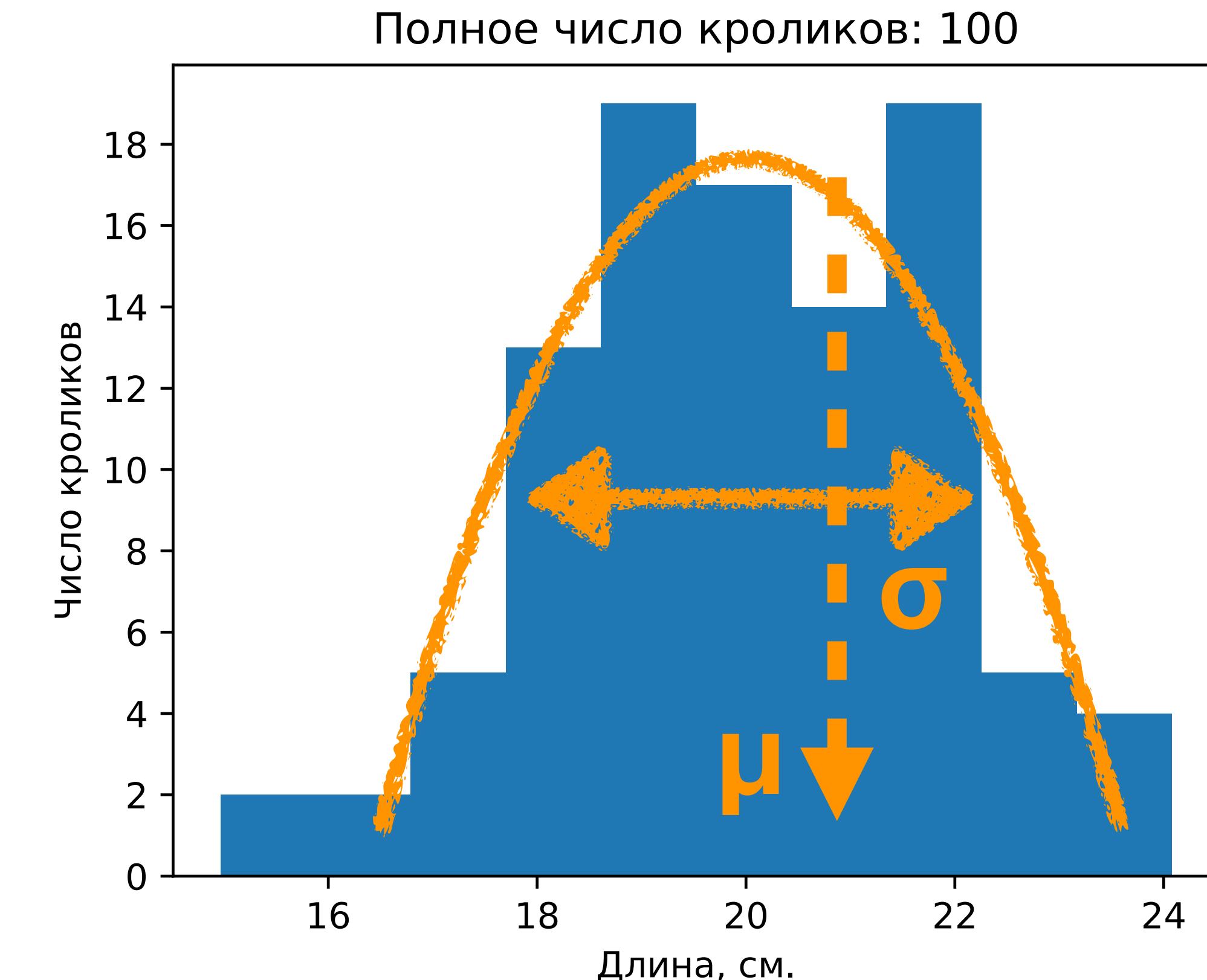
Оценочная статистика



Проверка гипотез



Правда ли, что вес зависит от длины линейно с гауссовым шумом?



Правда ли, что число кроликов зависит от длины нормально?

Python 3

Почему именно он?

- Хорошо спроектирован и интуитивен
- Используется повсеместно, даже если не считать ML
- Динамическая типизация (сокращает время написания кода)
- Не требует (AOT) компиляции, код — переносим
- Отличные математические библиотеки
 - Написаны на C и на Fortran
 - Продуманы возможности процессорной оптимизации (векторизация и т.п.)
 - Ваша программа на C/C++ практически наверняка окажется медленнее Python + Numpy

Альтернативы



- Не используется нигде, за исключением вычислений
- Проприетарный



- Мало где используется, за исключением статистических задач
- Широко используется в биоинформатике и финтехе
- В итоге всё равно придётся выучить

Julia — слишком молод, хотя и перспективен

JavaScript — спроектирован ужасно

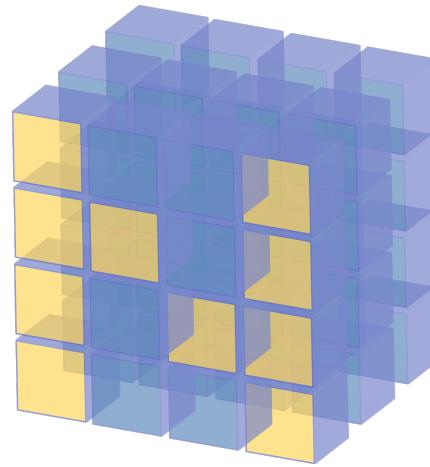


Jupyter Notebook

- Интерактивная среда поэтапного исполнения кода
- Запускается в браузере
- Позволяет поддерживать выполнение задач после его закрытия
- Ядра-интерпретаторы для языков Python, R, Julia, C++, Java...
- Удалённое управление серверами и кластерами через интернет-соединение

- Подсветка синтаксиса
- Поэтапное исполнение кода
- Использование т.н. magic'ов
 - %time line — измеряет время работы кода в строке
 - %%time — измеряет время работы кода в ячейке
 - %%timeit — делает то же самое, прогоняя код множество раз
 - %run file.py — выполняет питоновский файл
 - ! command — выполнение Shell-команд (в т.ч. внутри Python-кода)
 - object? — просмотр документации
 - %%python2, %%R, %%bash — выполнение кода других языков (если установлены ядра)
 - %magic — полный список команд. Также можно посмотреть по [ссылке](#)

IP[y]: IPython
Interactive Computing



NumPy



SciPy

- Работа с массивами произвольной размерности (ndarray)
- Линейная алгебра
- Генерация и работа со случайными числами
- Работа с распределениями случайных величин
- Численные методы

Другие библиотеки

- **Matplotlib** – построение графиков
- **Pandas** – работа с таблицами
- **Seaborn** – высокоуровневое построение графиков на базе Matplotlib
- **Statsmodels** – расширенный статистический пакет
- **Sklearn** – машинное обучение (за вычетом нейросетей)