Эксперименты

Лекция 6

Необходимость экспериментов

- Если X разработал программу P для решения задачи T, каким образом можно убедить других, что P действительно решает T?
- Иногда можно математически доказать, что P – корректная программа
- Однако во многих случаях это сделать слишком трудно или невозможно

Необходимость экспериментов

- Другая задача: доказать утверждение, что метод P превосходит метод Q
- Формально это сделать часто невозможно
  - Даже при анализе временной сложности в случае совпадения теоретических оценок
- Как правило, пользователям необходимо знать скорость работы в типичных случаях, а не в худшем случае
  - Формальные оценки часто делаются для худшего случая
- $\rightarrow$ необходимо экспериментальное исследование

Тестирование vs. оценка

- Тестирование (testing) – проверка соответствия программы спецификации
  - проверка *корректности* программы
- Программа должна выдавать результат для любого входного значения
- Результат должен соответствовать спецификации
- Невозможно проверить все возможные входные значения и пути выполнения программы
- Тестирование может доказать, что ошибки есть, но не может доказать, что ошибок нет

Тестирование vs. оценка

- Оценка (evaluation):
  - P – это более качественное решение, чем Q
    - Программа P сортирует числа быстрее, чем программа Q
  - P выдает результаты *хорошего качества*
    - Программа составления расписания доставки товаров создает расписание с минимально возможным временем доставки
  - Для типичных пользователей предпочтительнее использовать P, чем Q
    - Пользователи предпочитают браузер Chrome браузеру Edge

Тестирование vs. оценка

- Отличается от тестирования: качество vs. корректность
- Способы оценки:
  - провести множество запусков обеих программ сортировки на разных вариантах тестовых данных
  - сравнить результат работы программы составления расписания с известным наилучшим расписанием
  - провести опрос пользователей по предпочтениям относительно браузеров

Ключевые вопросы

- *Цель*: что исследование должно продемонстрировать?
- *Воспроизводимость* (reproducibility) – исследование можно повторить с получением тех же результатов?
- *Данные*: насколько адекватен выбор данных?
- *Анализ результатов*: объективно ли приведены и проанализированы результаты?
- *Масштабирование*: будет ли похожее поведение в случае использования более масштабных данных?

Цель

- Экспериментальное исследование должно предоставить доказательства, поддерживающие *гипотезу*
- Примеры:
  - данный алгоритм является эффективным
  - результаты выполнения алгоритма являются хорошим решением
  - данный алгоритм лучше альтернативных
- Гипотезу необходимо четко сформулировать!

Данные

- Варианты:
  - случайно сгенерированные данные
    - предположение: типичные данные = случайно сгенерированные данные
  - стандартные наборы данных (benchmarks)
    - UCI Machine Learning Repository (ссылка)
    - SuperGLUE Benchmark (ссылка)
    - Russian SuperGLUE (ссылка)
- Необходимо подробное описание данных

Критерии качества

- Необходимо определить, что значит «метод P **лучше** метода Q» $\rightarrow$ критерии качества

Критерии качества

- Например, для задач **классификации**:

		Оценка классификатора	
		$a(x) = +1$ (Positive)	$a(x) = -1$ (Negative)
Истинные ответы	$y = +1$	TP (True Positive)	FN (False Negative)
	$y = -1$	FP (False Positive)	TN (True Negative)

- Accuracy (Правильность - доля правильных ответов модели среди всех предсказаний):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- Precision (Точность - доля истинно положительных ответов среди всех положительных ответов модели):

$$P = \frac{TP}{TP + FP}$$

- Recall (Полнота - доля истинно положительных ответов среди всех правильных ответов):

$$R = \frac{TP}{TP + FN}$$

- F1-measure (F1-score, F1-мера - гармоническое среднее между точностью и полнотой):

$$F1 = \frac{2PR}{P + R}$$

Критерии качества

- для задач **регрессии**:

- **Mean Squared Error (MSE**) – среднее значение квадрата разности между предсказанными и правильными значениями.
- **Root Mean Squared Error (RMSE)** – квадратный корень из среднего значения квадратов разности между предсказанными и правильными значениями.
- **Mean Absolute Error (MAE)** – среднее значение абсолютной разности между предсказанными и правильными значениями.
- **R2-коэффициент детерминации** – мера, которая показывает, насколько хорошо модель подходит для данных. R2-коэффициент может принимать значения от 0 до 1, где 1 означает идеальное соответствие.

Критерии качества

- для задач **кластеризации**:

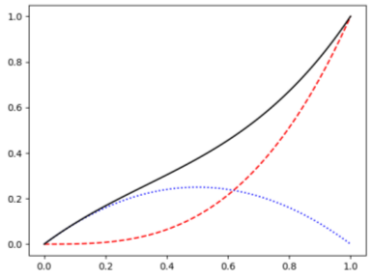
- **Silhouette score (Коэффициент силуэта)** – мера, которая показывает, насколько точно каждый объект соответствует своему кластеру и насколько он отличается от других кластеров. Значение коэффициента силуэта может варьироваться от -1 до 1, где 1 означает, что объекты внутри кластера находятся ближе друг к другу, чем к объектам других кластеров.

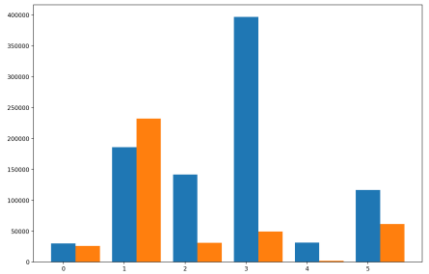
Анализ результатов

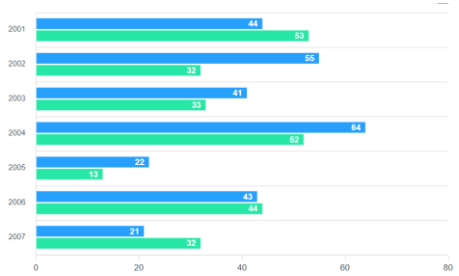
- Результаты поддерживают или опровергают гипотезу?
- «Да»:
  - чем это подтверждается?
  - как наиболее убедительно представить результаты?
  - достаточно ли проведено экспериментов?
- «Нет»:
  - результаты подтверждают альтернативную гипотезу?

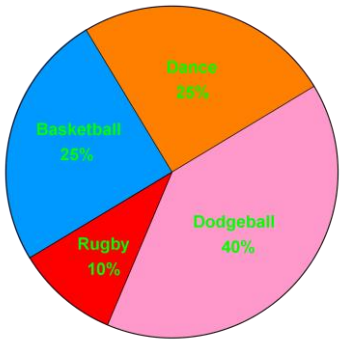
Представление результатов

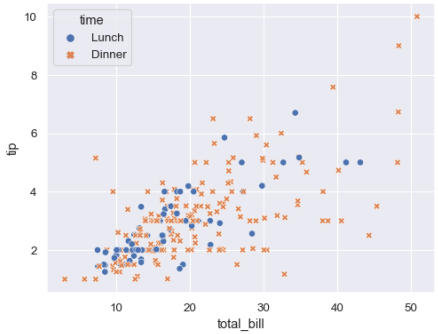
- Таблица
- График
- Столбчатая диаграмма (гистограмма)
- Круговая диаграмма
- Диаграмма рассеяния
- Ящик с усами

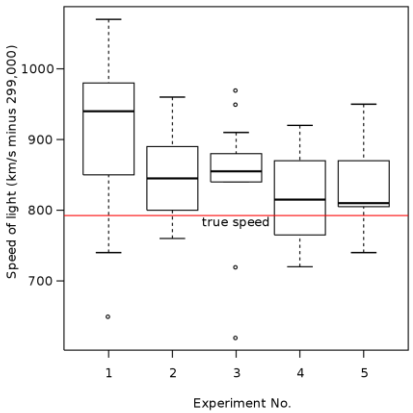






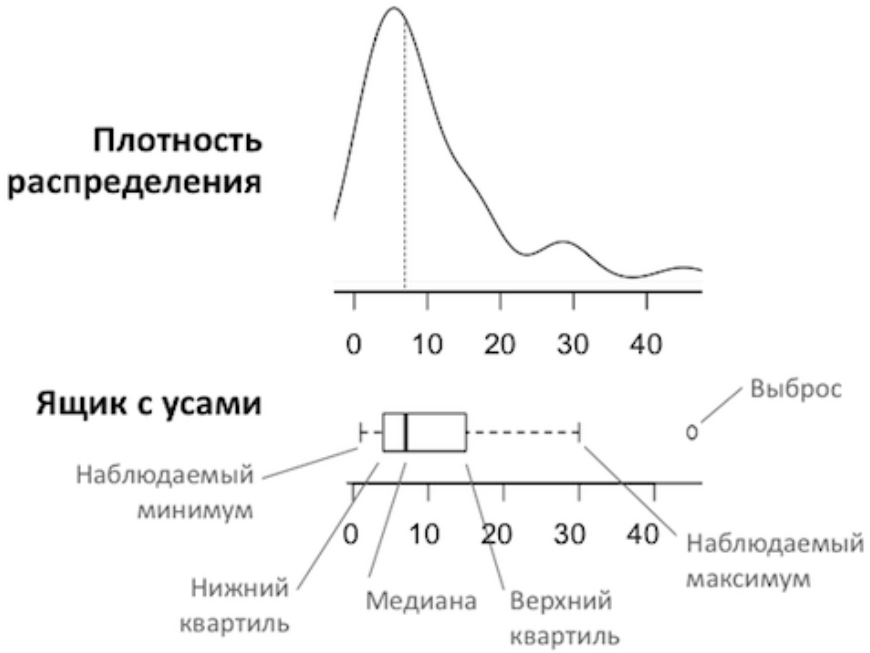




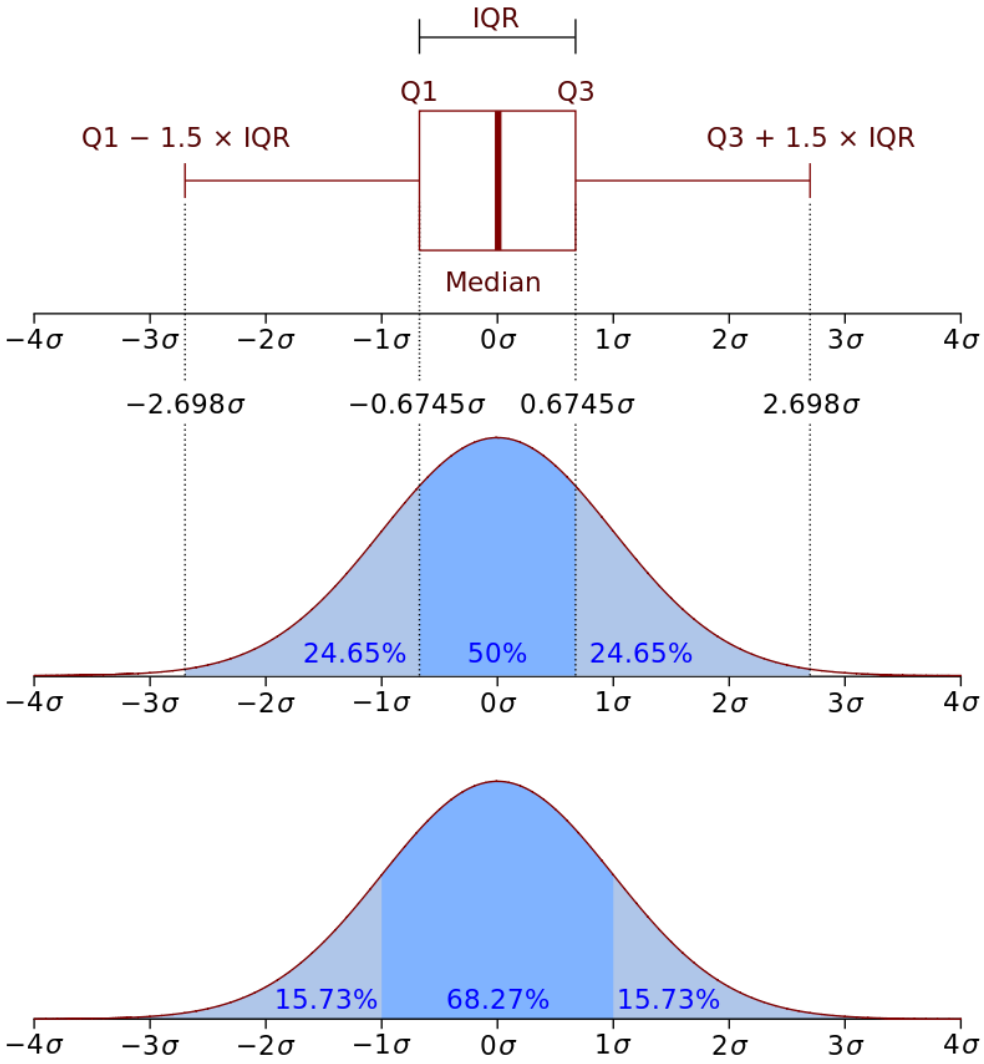


Представление результатов

- Параметры распределения:
  - среднее μ
  - медиана
  - мода
  - дисперсия σ^2
    - среднеквадратическое (стандартное) отклонение σ
  - минимум/максимум



Нормальное распределение



Эксперименты и обсуждение

План раздела «Эксперименты»:

- Описание данных
  - источники
  - описательные статистики
  - предобработка
- Методология проведения эксперимента
  - методы для сравнения (baselines)
  - критерии качества
- Реализация (инструменты)
  - параметры моделей/инструментов
- Результаты
  - таблицы и/или графики
  - выделение лучших
  - проверка статистической значимости
- Анализ результатов
  - часто объединяется с обсуждением результатов
  - анализ ошибок (примеры)
  - ablation study

Примеры

- Krishna et al. Reformulating Unsupervised Style Transfer as Paraphrase Generation (2020)
- Luong et al. Effective Approaches to Attention-based Neural Machine Translation (2015)
- Pennington et al. GloVe: Global Vectors for Word Representation (2014)
- Peyrard. A Simple Theoretical Model of Importance for Summarization (2019)
- Xu et al. Vocabulary Learning via Optimal Transport for Neural Machine Translation (2021)
- Logunov et al. Safety and efficacy of an rAd26 and rAd5 (2021)

Домашнее задание

- **Продумать и описать дизайн экспериментов:**
  - **Данные**
    - **источники**
    - **описательные статистики**
    - **предобработка**
  - **Методология**
    - **методы для сравнения (baselines)**
    - **критерии качества**
  - Инструменты
  - Представление результатов
  - Анализ результатов