ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Лекция 3

Частоты и группировка данных

Оспан Ә.Ғ., PhD, доцент-исследователь

КазНУ, кафедра ИИ и Big Data

2026-2027 учебный год

Цель и содержание лекции

Цель и результаты обучения

- Рассчитывать частоту, долю и процент с правильным знаменателем.
- Различать unique, nunique, value_counts и groupby.
- Проверять численности, границы интервалов и смысл групповых средних.

Содержание лекции

1. **Частоты**
2. **Уникальные значения**
3. **Доли**
4. **groupby**
5. **Интервалы**

Отдельная запись и группа: разная детализация

Наблюдение

- Одна строка исходника представляет запись об игроке.
- Признаки описывают конкретную запись.
- Фильтр выбирает часть исходных строк.

Группа

- Ключ объединяет строки с общим значением.
- Агрегат кратко описывает выбранные строки.
- При сжатии теряется часть индивидуальных различий.

После группировки изменяется смысл строки, поэтому его указывают заново.

Частота, доля и процент

$$share_c = n_c / N;\ percent_c = 100 * share_c$$

Абсолютная представленность

- n_c: число строк выбранной категории.
- Сумма частот равна числу учтенных строк.

Относительная представленность

- N задается вопросом и правилами учета.
- normalize=True дает долю, а не процент.

3 из 6 строк: доля 0.5, процент 50%.

value_counts: сколько записей каждого клуба?

Пример Python

```
import pandas as pd

demo = pd.DataFrame({
    'Club': ['A', 'A', 'B', 'B', 'B', 'C'],
    'Position': ['GK', 'ST', 'GK', 'ST', 'ST', 'CM'],
    'Age': [20, 24, 22, 26, 30, 21],
    'Wage': [10, 30, 20, 40, 60, 15]
})
counts = demo['Club'].value_counts()
print(counts.to_dict())
assert counts.sum() == len(demo)
```

Результат: {'B': 3, 'A': 2, 'C': 1}

Что считаем

- Каждое появление категории в столбце.
- Индекс результата содержит категории.

Что проверяем

- Сумму счетчиков и все ничьи лидеров.
- Единицу строки: запись, а не полный состав.

unique, nunique и value_counts отвечают по-разному

Вопрос	Инструмент	Результат
Какие категории встретились?	unique	Значения в порядке появления
Сколько разных категорий?	nunique	Одно число
Сколько строк каждой категории?	value_counts	Series счетчиков
Каков показатель каждой группы?	groupby + агрегация	Series или DataFrame
Сколько людей?	Надежный идентификатор	Имени недостаточно

Знаменатель и пропуски меняют долю

Полный файл

- Числитель: записи нужной категории.
- Знаменатель: все учитываемые строки файла.
- Сумма долей должна соответствовать правилам учета.

Подвыборка

- Сначала выбираем нужную совокупность.
- Знаменатель: число ее учтенных записей.
- Пропуски включаем или исключаем явно.

Одна частота может давать разные доли при разных явно заданных знаменателях.

Из Series частот в таблицу отчета

Пример Python

```
frequency = (
    counts.rename_axis('Club')
    .reset_index(name='n_players')
)
frequency['share'] = frequency['n_players'] / len(demo)
frequency['percent'] = 100 * frequency['share']
print(frequency)
assert frequency.shape == (3, 4)
```

Результат: B: 3 и 50%; A: 2 и 33.33%; C: 1 и 16.67%.

Явные поля

- Club: категория; n_players: число строк.
- reset_index превращает метки в обычный столбец.

Шкалы

- share измеряется в долях единицы.
- percent = 100 * share.

groupby: разделить, вычислить, собрать

1

Ключ группы

Club или комбинация Club и Position.

2

Показатель

Age, Wage либо другой содержательно выбранный признак.

3

Функция

Среднее, сумма, минимум или счетчик по смыслу вопроса.

4

Контроль

Число групп, численности и независимая проверка ячейки.

- Объект группировки задает разбиение, а не готовый ответ.
- Поля выбираем явно: числовой тип не делает сумму осмысленной.

Средний возраст, размер и сумма по клубам

Пример Python

```
grouped = demo.groupby('Club', sort=True)
print(grouped['Age'].mean().to_dict())
print(grouped.size().to_dict())
print(grouped['Wage'].sum().to_dict())
```

Результат: Age: A=22, B=26, C=21. Размеры: 2, 3, 1. Wage: 40, 120, 15.

Одна группировка

- Каждый ключ Club определяет набор строк.
- Результат меняется с выбранным агрегатом.

Три вопроса

- Средний возраст; число записей; сумма Wage.
- Проверяем небольшой клуб вручную.

size и count совпадают не всегда

Метод	Что считает	Если один Wage неизвестен
size	Все строки группы	Запись остается в счетчике
count по Wage	Заполненные Wage	Счетчик уменьшается
nunique по Club	Разные значения Club	Не является числом записей
mean по Wage	Среднее известных значений	Меняется используемая совокупность

Показывайте численность группы и число значений, участвующих в показателе.

Числовые интервалы: ширина и границы

Пример Python

```
auto = demo['Age'].value_counts(bins=3, sort=False)
print(auto)
age_group = pd.cut(
    demo['Age'], bins=[17, 21, 25, 30],
    labels=['18-21', '22-25', '26-30'],
    include_lowest=True
)
print(age_group.value_counts(sort=False).to_dict())
assert age_group.notna().all()
```

Результат: {'18-21': 2, '22-25': 2, '26-30': 2}

Автоматические интервалы

- bins задает равную ширину диапазона.
- Численности интервалов могут различаться.

Объявленные границы

- cut фиксирует сопоставимые диапазоны.
- Проверяем включенность и выход за пределы.

Общее среднее требует размеров групп

overall_mean = sum(n_g * mean_g) / sum(n_g)

Правильный вес

- n_g: число наблюдений в групповом среднем.
- Большая группа вносит больше исходных значений.

Другой вопрос

- Простое среднее групп дает равный вес группам.
- При разных размерах оно обычно отличается.

(2*22 + 3*26 + 1*21) / 6 = 23.8333; (22+26+21)/3 = 23.

Группы football: размер и пределы интерпретации

Размер

- Среднее двух записей чувствительно к одному игроку.
- Рядом с агрегатом показываем численность.

Охват

- Представленность клуба относится к файлу.
- Исторический срез не равен текущему составу.

Шкалы

- Wage/Value: единицы источника.
- Penalties: рейтинг, не голы.

Высокое среднее небольшой группы требует проверки ее состава и численности.

Частые ошибки группового анализа

Счетчик

- nunique не равно числу строк.
- Все ничьи минимума и максимума сохраняем.

Доля

- Доля и процент используют разные шкалы.
- Знаменатель задает вопрос, а не удобство.

Сжатие

- bins означает ширину, а не равную численность.
- Средние объединяем с обоснованными весами.

Перед выводом называем единицу строки, знаменатель и контрольный итог таблицы.

Пять опор группового отчета

1. Частота считает записи; доля делит их на объявленную совокупность.
2. unique, nunique и value_counts дают разные типы ответа.
3. groupby задает разбиение, агрегат формирует показатель группы.
4. size считает строки, count считает заполненные значения.
5. Интервалы и групповые средние требуют контроля границ и весов.

Проверьте знаменатели и группы

1. Чем частота отличается от доли и процента?
2. Когда count дает меньшее значение, чем size?
3. Почему unique по Name не устанавливает число людей?
4. Что гарантирует bins=4, а чего он не гарантирует?
5. Почему среднее групповых средних может отличаться от общего?

Самостоятельная работа

Добавьте один пропуск Wage в demo. Сравните size, count и mean по клубам; затем восстановите общее среднее с правильными весами.

Источники по частотам и группировке

Документация

1. **pandas: группировка**
   https://pandas.pydata.org/docs/user_guide/groupby.html
2. **pandas: value_counts**
   https://pandas.pydata.org/docs/reference/api/pandas.Series.value_counts.html
3. **pandas: интервалы cut**
   https://pandas.pydata.org/docs/reference/api/pandas.cut.html

Основа лекции

- Лекция 3.pdf
- Лабораторная работа 3.docx
- Лабораторная работа 3 с ответами.docx