ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Лекция 11

Вероятность и описательная статистика

Оспан Ә.Ғ., PhD, доцент-исследователь

КазНУ, кафедра ИИ и Big Data

2026-2027 учебный год

Цель и содержание лекции

Цель и результаты обучения

- Отличать вероятность модели от наблюдаемой частоты
- Рассчитывать ожидание, дисперсию и стандартное отклонение
- Различать разброс наблюдений и разброс выборочных средних

Содержание лекции

1. **Случайный эксперимент**
2. **Распределения**
3. **Ожидание и дисперсия**
4. **Нормальная модель**
5. **Статистика и ЦПТ**

Вероятность модели и частота в опыте

До эксперимента

- Вероятность описывает выбранную модель опыта.
- Для честного кубика P(четное) = 3/6 = 0.5.
- Равновероятность исходов должна быть обоснована.

После серии опытов

- Частота зависит от реально полученных результатов.
- 12 четных исходов из 20 дают частоту 0.6.
- Конечная серия не обязана точно повторять модель.

Сначала определите эксперимент, событие и механизм отбора.

Способ выбора меняет веса наблюдений

Эксперимент	P(X=10)	P(X=40)	E(X)
Выбор одного из трех курсов	2/3	1/3	20
Выбор одного из 60 студентов	1/3	2/3	30
Различие в вопросе	Вес по курсам	Вес по людям	Разные единицы анализа

Большая группа чаще встречается при выборе ее участника.

Случайная величина и ее распределение

Дискретная модель

- Конечное или счетное множество значений.
- Вероятности p_i неотрицательны; их сумма равна 1.
- Пример: число купленных книг.

Непрерывная модель

- Вероятности относятся к интервалам значений.
- Плотность задает вероятность через площадь.
- Пример: время ожидания транспорта.

Числовая запись категории сама по себе не дает количественную шкалу.

Математическое ожидание - взвешенный центр

$$E(X) = \sum(x_i * p_i)$$

Что суммируем

- Каждое допустимое значение x_i с весом p_i.
- Частый исход сильнее влияет на взвешенный центр.

Как интерпретируем

- Средний результат повторений заданной модели.
- Ожидание не обязано быть допустимым исходом.

X: 1, 2, 3; p: 0.3, 0.4, 0.3. E(X)=2.

Проверяем распределение и вычисляем его центр

Пример Python

```
import numpy as np

x = np.array([1., 2., 3.])
p = np.array([0.3, 0.4, 0.3])
assert x.shape == p.shape
assert np.all(p >= 0) and np.isclose(p.sum(), 1)
mu = np.sum(x * p)
var = np.sum((x - mu) ** 2 * p)
print(mu, round(var, 3))
```

Результат: 2.0 0.6

Проверка условий

- Массивы имеют одинаковую форму.
- Вероятности неотрицательны и дают единицу.

Детерминированный расчет

- mu - центр заданного распределения.
- var - взвешенный квадрат отклонений.

Дисперсия и стандартное отклонение

$$\mathrm{Var}(X) = E((X - E(X))^2); \quad \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Дисперсия

- Отклонения возводятся в квадрат и усредняются.
- Измеряется в квадрате исходной единицы.
- Всегда неотрицательна; у константы равна нулю.

Стандартное отклонение

- Квадратный корень из дисперсии.
- Имеет ту же единицу, что и X.
- Не равно среднему абсолютному отклонению.

Для числа книг: Var(X)=0.6; SD(X) примерно 0.7746.

Преобразования: центр и разброс меняются по-разному

Преобразование	Ожидание	Дисперсия
Y=aX+b	aE(X)+b	a**2 Var(X)
X+Y: общий случай	E(X)+E(Y)	Var(X)+Var(Y)+2Cov(X,Y)
X+Y: независимые	E(X)+E(Y)	Var(X)+Var(Y)
Константа c	c	0

Линейность ожидания и сложение дисперсий имеют разные условия.

Плотность и накопленная вероятность

Плотность f(x)

- Высота кривой не является вероятностью точки.
- Вероятность интервала равна площади под кривой.
- Вся площадь равна 1; f(x) может быть выше 1.

Функция распределения F(x)

- F(x)=P(X<=x) накапливает вероятность слева.
- P(a<X<=b)=F(b)-F(a).
- ppf находит границу, sf - правый хвост.

Для гистограммы плотности проверяют площади, а не сумму высот.

Нормальная модель: центр и масштаб



Как читать нормальную плотность

- mu определяет центр симметрии, sigma>0 - масштаб.
- z=(x-mu)/sigma измеряет отклонение в единицах SD.
- Стандартизация не превращает любую форму в нормальную.
- Правило трех сигм требует нормальной модели.

Нормальность - предположение о модели, которое требует проверки.

Нормальные хвосты и квантили в SciPy

Пример Python

```
from scipy.stats import norm

upper = norm.sf(190, loc=182, scale=4)
middle = norm.cdf(186, loc=182, scale=4)
middle -= norm.cdf(178, loc=182, scale=4)
boundary = norm.ppf(0.975, loc=182, scale=4)
print(round(upper, 4), round(middle, 4))
print(round(boundary, 2))
```

Результат: 0.0228 0.6827; квантиль 189.84

Прямая задача

- sf: вероятность значения выше 190.
- Разность cdf: вероятность интервала 178-186.

Обратная задача

- ppf: граница для накопленной доли 0.975.
- loc - среднее; scale - стандартное отклонение.

Описание набора и оценивание распределения

Конечный наблюдаемый набор

- Среднее, медиана, Q1, Q3 и IQR=Q3-Q1.
- Для описательной дисперсии возможен ddof=0.
- Число описывает именно имеющиеся строки.

Оценка параметра по выборке

- Обычно s**2 делит сумму квадратов на n-1.
- NumPy и pandas имеют разные defaults ddof.
- Отбор и независимость требуют отдельного обоснования.

В football.csv Penalties - рейтинг навыка, а не число голов.

ЦПТ: распределение среднего, а не исходных данных

E(X_bar)=mu; SE(X_bar)=sigma/sqrt(n)

Что меняется при повторных выборках

- Каждая выборка дает свое среднее X_bar.
- Стандартная ошибка описывает разброс этих средних.

Условия и ограничения

- Независимость, одинаковый механизм, конечная дисперсия.
- Нет универсального правила «n>30 достаточно».
- ЦПТ не делает исходные значения нормальными.

Увеличение n в 4 раза уменьшает стандартную ошибку в 2 раза.

Ошибки при переходе от чисел к выводам

Модель

- Не считать любой список исходов равновероятным.
- Не складывать дисперсии зависимых навыков.

Характеристика

- Не путать плотность и вероятность точки.
- Не путать SD наблюдений и SE среднего.

Границы данных

- Не переносить описание файла на всех спортсменов.
- Не считать большой n доказательством независимости.

Главные результаты лекции

1. Вероятность задает модель; частота описывает конечную серию.
2. Способ выбора объекта определяет веса и распределение.
3. Ожидание - взвешенный центр; дисперсия - средний квадрат отклонения.
4. Непрерывные вероятности определяются площадями и cdf.
5. Стандартная ошибка среднего уменьшается как 1/sqrt(n).

Проверка понимания и самостоятельная работа

1. Почему E(X) может не быть допустимым исходом?
2. Чем выбор клуба отличается от выбора клуба случайного игрока?
3. Когда можно сложить дисперсии без ковариации?
4. Почему стандартизация не гарантирует нормальность?
5. Что различает SD наблюдений и SE среднего?

Самостоятельная работа

Для X=0,2,5 с вероятностями 0.2,0.5,0.3 найдите E(X), Var(X) и характеристики Y=3X-4. Проверьте прямым расчетом.

Документация и исходные материалы

Документация

1. NumPy: генераторы случайных чисел

https://numpy.org/doc/stable/reference/random/generator.html

2. SciPy: нормальное распределение

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html

3. pandas: описательная статистика

https://pandas.pydata.org/docs/user_guide/basics.html#descriptive-statistics

Основа лекции

- Лекция 11 ИАД рус.pptx
- Лабораторная работа 11.docx
- Лабораторная работа 11 с ответами.docx