ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Лекция 12

Доверительные интервалы

Оспан Ә.Ғ., PhD, доцент-исследователь

КазНУ, кафедра ИИ и Big Data

2026-2027 учебный год

Цель и содержание лекции

Цель и результаты обучения

- Различать параметр, точечную оценку и стандартную ошибку
- Обосновать z-, t- и bootstrap-интервал
- Объяснять уровень доверия через повторение процедуры

Содержание лекции

1. **Параметр и оценка**
2. **Покрытие интервалов**
3. **Среднее**
4. **Доля**
5. **Bootstrap и ограничения**

Параметр фиксирован, оценка зависит от выборки

Параметр распределения

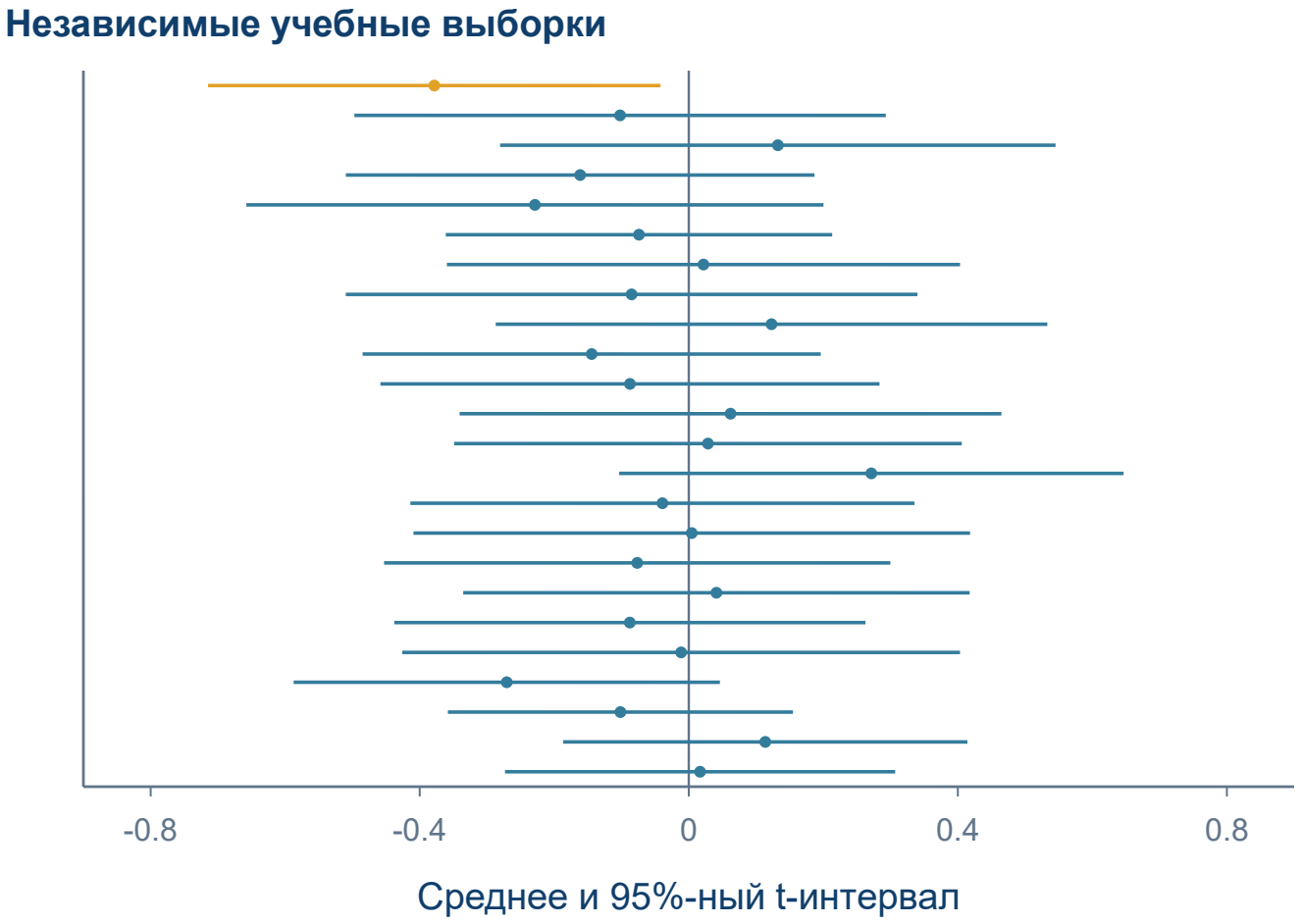
- mu, sigma и p фиксированы в принятой модели.
- Обычно неизвестны и относятся к целевой совокупности.
- Их смысл определяется вопросом и способом отбора.

Статистика выборки

- Среднее, s и p_hat вычисляются по наблюдениям.
- Меняются при повторном получении выборки.
- Точечная оценка не показывает всю неопределенность.

Интервал не исправляет смещенный отбор или неверную шкалу.

Что означает доверительный уровень 95%



Повторяется вся процедура

- Каждая новая выборка дает новые границы L и U.
- При выполнении модели около 95% интервалов накрывают mu.
- После расчета конкретный интервал либо содержит mu, либо нет.
- Это не интервал для 95% отдельных наблюдений.

Доверие относится к свойству процедуры при ее предпосылках.

Стандартная ошибка среднего

$$SE(\bar{X}) = \sigma/\sqrt{n}$$

Разные меры разброса

- sigma - разброс отдельных значений.
- SE - разброс средних при повторных выборках.
- При неизвестной sigma используют оценку s/sqrt(n).

Влияние объема

- Увеличение n в 4 раза уменьшает SE в 2 раза.
- Требуется независимый одинаковый механизм наблюдений.
- Большое n не устраняет систематический сдвиг.

sigma=10, n=100: SE=1; для n=400: SE=0.5.

Z-интервал: sigma действительно известна

$$\bar{X} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Предпосылки

- Независимая случайная выборка.
- Нормальная модель среднего: точная или обоснованная.
- sigma известна из внешней надежной информации.

Два хвоста

- confidence=0.95 означает alpha=0.05.
- Критический квантиль: norm.ppf(0.975).
- Полуширина равна критическому значению, умноженному на SE.

X_bar=50, sigma=10, n=100: интервал [48.04; 51.96].

Сравниваем Z- и T-интервалы на заданных условиях

Пример Python

```
import numpy as np
from scipy.stats import norm, t

z_half = norm.ppf(0.975) * 10 / np.sqrt(100)
t_half = t.ppf(0.975, df=15) * 8 / np.sqrt(16)
print(round(50-z_half, 2), round(50+z_half, 2))
print(round(50-t_half, 2), round(50+t_half, 2))
```

Результат: Z: 48.04 51.96; T: 45.74 54.26

Известная sigma

- Нормальное критическое значение.
- n=100; sigma=10; среднее 50.

Неизвестная sigma

- t-квантиль с 15 степенями свободы.
- n=16; s=8; среднее 50.

Почему при неизвестной sigma появляется T

Нормальная независимая выборка

- (X_bar-mu)/(s/sqrt(n)) имеет t-распределение.
- Степени свободы: n-1; s считают с ddof=1.
- Тяжелые хвосты учитывают оценивание sigma.

Что T не исправляет

- Систематическое смещение отбора.
- Зависимые повторные наблюдения.
- Сильную асимметрию маленькой выборки без анализа.

При большом n t и z близки, но предпосылки остаются важны.

Метод выбирают после определения дизайна

Оцениваемый объект	Основа	Ключевая проверка
Среднее, sigma известна	z * sigma/sqrt(n)	Известна ли sigma вне этой выборки?
Среднее, sigma неизвестна	t * s/sqrt(n)	Дизайн и форма распределения
Доля	Вальд или Уилсон	Редкие события, независимость
Сложная статистика	Bootstrap-квантили	Правильная единица пересэмплирования

Метод не устраняет неизвестный отбор и зависимость наблюдений.

Интервал доли: удобство и пределы Вальда

$$p_hat +/- z * sqrt(p_hat*(1-p_hat)/n)$$

Простой интервал Вальда

- p_hat=k/n; каждое наблюдение - независимый исход 0/1.
- Нужны достаточные числа успехов и неуспехов.
- Вблизи 0 и 1 нормальное приближение ненадежно.

Альтернатива: Уилсон

- Корректирует центр и полуширину.
- Границы остаются в диапазоне [0;1].
- Не исправляет добровольный самовыбор ответивших.

Границы метода и ограничения отбора - разные проблемы.

Вычисляем интервал Уилсона для 60 успехов из 100

Пример Python

```
k, n = 60, 100
z = norm.ppf(0.975)
phat = k / n
d = 1 + z**2 / n
c = (phat + z**2 / (2*n)) / d
term = phat*(1-phat)/n + z**2/(4*n**2)
h = z * np.sqrt(term) / d
print(round(c-h, 3), round(c+h, 3))
```

Результат: 0.502 0.691

Входные условия

- 60 успехов среди 100 независимых наблюдений.
- Двусторонний доверительный уровень 95%.

Чтение результата

- Интервал оценивает долю процесса.
- Уровень доверия относится к процедуре.

Симуляция показывает смысл покрытия

1

Задать известный mu

Нормальная модель: mu=50, sigma=10.

2

Повторить отбор

Получить много независимых выборок одного размера.

3

Построить интервалы

Применить одну заранее выбранную процедуру.

4

Посчитать покрытия

Проверить L<=50<=U в каждом повторе.

- Уровень доверия относится к серии процедур.
- Доля покрытий в конечной симуляции не обязана быть ровно 0.95.

Bootstrap: повторный отбор из наблюдаемого набора

1

Наблюдаемый набор

Определить статистику и независимую единицу отбора.

2

Выборка с возвращением

Повтор имеет исходный размер; строки могут повторяться.

3

Повторный расчет

Сохранить среднее или другую выбранную статистику.

4

Границы по квантилям

Для 95% взять квантили 0.025 и 0.975.

- Обычный bootstrap по строкам предполагает независимость.
- Малый набор не содержит неизвестных редких исходов.

Учебный bootstrap среднего возраста

Пример Python

```
from pathlib import Path
import pandas as pd

football = pd.read_csv(Path('data') / 'football.csv')
age = football['Age'].to_numpy(dtype=float)
rng = np.random.default_rng(42)
means = np.empty(1000)
for i in range(means.size):
    sample = rng.choice(age, size=age.size, replace=True)
    means[i] = sample.mean()
print('Среднее файла:', age.mean())
print(np.quantile(means, [0.025, 0.975]))
```

Что делает код

- 1000 повторов с возвращением по игрокам.
- Границы - квантили средних 0.025 и 0.975.

Как ограничить вывод

- Независимость строк является допущением.
- Клубные связи и неизвестный отбор не учтены.

Среднее файла известно; неопределенность относится к принятой модели.

Что доверительный интервал не гарантирует

Интерпретация

- Не вероятность mu после вычисления границ.
- Не вероятность правильности одного вывода.

Единица анализа

- Не диапазон 95% отдельных значений.
- Не автоматическая оценка всех будущих наблюдений.

Дизайн данных

- Не исправляет смещение и зависимость.
- Не определяет неизвестную генеральную совокупность.

Главные результаты лекции

1. Параметр фиксирован; статистики и границы меняются между выборками.
2. 95% доверия - свойство повторяемой процедуры покрытия.
3. Известная sigma ведет к z; оцененная s - к t при нужной модели.
4. Интервал доли Вальда ненадежен при редких событиях.
5. Bootstrap требует правильной единицы и схемы пересэмплирования.

Проверка понимания и самостоятельная работа

1. Что случайно до получения данных при построении интервала?
2. Почему интервал среднего не является диапазоном 95% измерений?
3. Чем известная sigma отличается от выборочного s?
4. Почему обрезание границ Вальда не исправляет покрытие?
5. Что ограничивает bootstrap по игрокам одного клуба?

Самостоятельная работа

При n=36, X_bar=120 и известной sigma=18 сравните интервалы 90%, 95% и 99%. Объясните, почему меняется ширина.

Документация и исходные материалы

Документация

1. SciPy: распределение Стьюдента

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

2. SciPy: bootstrap

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

3. NumPy: quantile

https://numpy.org/doc/stable/reference/generated/numpy.quantile.html

Основа лекции

- Лекция 12 Доверительные интервалы на Python.pptx
- Лабораторная работа 12 Доверительные интервалы.docx
- Лабораторная работа 12 с ответами Доверительные интервалы.docx