**Лекция 12**

**Центральные предельные теоремы**

*(Почему нормальное распределение правит миром данных)*

**1. Что говорит ЦПТ?**

**Центральная предельная теорема (ЦПТ)** утверждает:

Если брать **большое количество выборок** (например, 1000 раз по 30 элементов) из любого распределения (даже ненормального!), то распределение **средних значений этих выборок** будет стремиться к нормальному распределению.\*\*

**Почему это важно?**

* Объясняет, почему нормальное распределение так распространено в природе и данных.
* Позволяет использовать методы статистики даже для ненормальных исходных данных.

**2. Формулировка теоремы**

Пусть X1,X2,...,Xn — независимые, одинаково распределённые случайные величины с:

* Математическим ожиданием μ.
* Конечной дисперсией σ^2.

Тогда при n→∞ распределение выборочного среднего X=X1+X2+...+Xn / n​​ будет стремиться к нормальному распределению:

X∼N(μ,σ^2 / n).

**Простыми словами:**

* Средние значения выборок будут группироваться вокруг истинного среднего μ.
* Разброс средних уменьшается с ростом n

**3. Демонстрация на Python**

**Задача:** Показать, как средние выборок из равномерного распределения становятся нормальными.

**Шаги:**

1. Сгенерируем 1000 выборок по 30 элементов из равномерного распределения.
2. Для каждой выборки вычислим среднее.
3. Построим гистограмму средних.

**Код:**

import numpy as np

import matplotlib.pyplot as plt

# Параметры

n\_samples = 1000 # Количество выборок

sample\_size = 30 # Размер каждой выборки

data = []

# Генерация данных

for \_ in range(n\_samples):

sample = np.random.uniform(0, 1, sample\_size) # Равномерное распределение

data.append(np.mean(sample))

# Визуализация

plt.hist(data, bins=30, density=True, alpha=0.6, label='Средние выборок')

# Нормальное распределение для сравнения

mu = 0.5 # Среднее равномерного распределения [0,1]

sigma = np.sqrt(1/12 / sample\_size) # Дисперсия равномерного: σ² = (b-a)²/12

x = np.linspace(0.3, 0.7, 100)

plt.plot(x, 1/(sigma \* np.sqrt(2 \* np.pi)) \* np.exp(-0.5 \* ((x - mu)/sigma)\*\*2), 'r-', lw=2, label='Теория')

plt.title("Распределение средних выборок")

plt.legend()

plt.show()

**Результат:**

* Гистограмма средних будет напоминать колокол Гаусса, даже если исходные данные равномерные!

**4. Условия применимости ЦПТ**

1. **Независимость:** Элементы выборки не должны влиять друг на друга.
2. **Одинаковое распределение:** Все данные взяты из одной генеральной совокупности.
3. **Конечная дисперсия:** Дисперсия не должна быть бесконечной (например, распределение Коши не подходит).

**Что делать, если условия нарушены?**

* Увеличить размер выборки n.
* Использовать методы, устойчивые к ненормальности (например, бутстреп).

**5. Практические примеры**

**5.1. Оценка среднего дохода**

**Задача:** Оценить средний доход в городе, где распределение доходов сильно скошено (богатые люди — редки, но их доходы огромны).

**Решение:**

* Собрать множество случайных выборок по 100 человек.
* По ЦПТ средние этих выборок будут нормально распределены.
* Построить доверительный интервал для среднего.

**5.2. A/B-тестирование**

**Задача:** Сравнить конверсию двух версий сайта.

**Как помогает ЦПТ?**

* Конверсия — биномиальное распределение (пользователь либо купил, либо нет).
* При большом трафике распределение средней конверсии становится нормальным.
* Можно применять t-тесты для сравнения.

**6. Ограничения ЦПТ**

1. **Медленная сходимость:** Для сильно скошенных распределений (например, экспоненциального) может потребоваться очень большой n.
2. **Не работает для «тяжёлых хвостов»:** Если данные имеют выбросы с бесконечной дисперсией (распределение Коши).

**Пример с распределением Коши:**

# Генерация данных из распределения Коши

data\_cauchy = np.random.standard\_cauchy(1000)

plt.hist(data\_cauchy, bins=50, range=(-10, 10), density=True)

plt.title("Распределение Коши")

plt.show()

Даже при большом n средние не будут нормальными!