

## Выборка. Порядковая статистика.

**Выборка** или **выборочная совокупность** — часть **генеральной совокупности** элементов, которая охватывается экспериментом (наблюдением, опросом).

Характеристики выборки:

- Качественная характеристика выборки — что именно мы выбираем и какие способы построения выборки мы для этого используем.
- Количественная характеристика выборки — сколько случаев выбираем, другими словами объём выборки.

А теперь важное: Выборка в математической статистике

Последовательность независимых случайных величин  $x_1, x_2, \dots, x_n$ , соответствующих всем возможным результатам  $n$  статистических экспериментов и имеющих одинаковый закон распределения вероятностей со случайной величиной  $X$ , называется выборкой объёма  $n$ , порождённой случайной величиной  $X$ . Если  $X$  — дискретная случайная величина, то выборкой объёма  $n$  называется любое подмножество  $n$  объектов генеральной совокупности объёма  $N$ , выбранное равновероятно среди всех таких подмножеств.

Буква  **$X$  (большая)** означает, что это **случайная величина**.

- В теории вероятностей и статистике **случайные величины** обозначаются большими буквами (например,  $X, Y, Z$ ), потому что их значения заранее неизвестны.
- Когда мы говорим о **выборке**, мы рассматриваем несколько случайных величин **одновременно**.
  - Если у нас есть выборка из  $n$  наблюдений, то у нас  $n$  случайных величин:
$$X_1, X_2, \dots, X_n$$
- Это означает, что **каждое  $X_i$  (где  $i = 1, \dots, n$ ) — это случайная величина, полученная из одного и того же распределения  $F(x)$** .

Тогда может возникнуть следующий вопрос: Почему выборка состоит из нескольких случайных величин?

Потому что в реальности мы **не можем наблюдать всё распределение**, а имеем только ограниченное количество данных.

Пример:

- Допустим, мы изучаем рост людей. Рост — случайная величина  $X$ , имеющая распределение (например, нормальное).
- Мы измеряем рост у **разных людей** — и каждое измерение даёт нам случайную величину:

$$X_1 = 170, X_2 = 165, X_3 = 180, \dots, X_n = 175$$

Таким образом, **каждый  $X_i$  — это реализация одной и той же случайной величины  $X$ , но на разных наблюдениях.**

**Почему говорят, что  $X_1, X_2, \dots, X_n$  распределены по  $F(x)$ ?**

Это означает, что:

1. Все элементы выборки **независимы** (обычно предполагается независимая выборка).
2. Все элементы выборки **имеют одно и то же распределение  $F(x)$**  (например, нормальное, экспоненциальное и т. д.).

#### **Пример:**

- Если мы бросаем монету 10 раз, то каждый раз исход может быть **Орел** или **Решка**.
- Тогда каждая случайная величина  $X_i$  (где  $X_1 = 1$ , если орёл, и  $X_i = 0$ , если решка) **имеет одно и то же распределение** — биномиальное с параметром  $p$ .

**В чем разница между  $X_i$  и конкретными значениями?**


- $X_i$  — это случайная величина (до эксперимента мы не знаем, чему она равна).
- $x_i$  — это конкретное значение, которое получилось в эксперименте.

#### **Пример:**

- Пусть  $X_1, X_2, \dots, X_5$  — это выборка случайных величин, обозначающих оценки студентов.
- Мы не знаем заранее их значения, но после эксперимента получаем:

$$x_1 = 80, x_2 = 75, x_3 = 90, x_4 = 85, x_5 = 70$$

- Здесь  $x_1, x_2, \dots, x_5$  — это конкретные числа, которые мы наблюдаем.

 **Важно:** выборка может быть **разной каждый раз**, но она подчиняется одному и тому же распределению  $F(x)$ .

### Вывод:

- Выборка состоит из случайных величин, потому что каждое наблюдение — это случайная величина.
- Все элементы выборки  $X_1, \dots, X_n$  имеют одно и то же распределение  $F(x)$ .
- До наблюдения каждое  $X_i$  — случайная величина, а после наблюдения — конкретное значение  $x_i$ .

### Порядковые статистики

Допустим, у нас есть выборка случайных величин:

$$X_1, X_2, \dots, X_n$$

Все  $X_i$  независимы и имеют одно и то же распределение  $F(x)$ .

Если мы отсортируем элементы выборки по возрастанию, получим **упорядоченную выборку**:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

Где:

- $X_{(1)}$  — **минимальное значение выборки** (первая порядковая статистика)
- $X_{(n)}$  — **максимальное значение выборки** (n-я порядковая статистика)
- $X_{(k)}$  — **k-я порядковая статистика**, то есть **k-й по величине элемент**

### 💡 Пример:

Допустим, у нас есть выборка из пяти случайных чисел:

$$X_1 = 7.1, X_2 = 5.3, X_3 = 6.4, X_4 = 9.2, X_5 = 8.0$$

После сортировки по возрастанию получаем:

$$X_{(1)} = 5.3, \quad X_{(2)} = 6.4, \quad X_{(3)} = 7.1, \quad X_{(4)} = 8.0, \quad X_{(5)} = 9.2$$

Здесь:

- **$X_{(1)}=5.3$  — минимум**
- **$X_{(5)}=9.2$  — максимум**
- **$X_{(3)}=7.1$  — медиана (если n нечётное)**

## Распределение порядковой статистики

Как найти  $F_{(k)}(x)$  — функцию распределения  $k$ -й порядковой статистики:

- Функция распределения  **$k$ -й порядковой статистики** получается через биномиальные вероятности:

$$F_{(k)}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^n F(x)^j \cdot (1 - F(x))^{n-j} \cdot C_n^j$$

- Для экстремальных значений формулы упрощаются:

- Для **минимума**  $X_{(1)}$ :

$$F_{(1)}(x) = 1 - (1 - F(x))^n$$

- Для **максимума**

$$F_{(n)}(x) = F(x)^n$$

Таким образом функции плотностей распределения будут выглядеть:

$$f(x) = \frac{dF(x)}{dx}$$

Тогда:

- Для **минимума**  $X_{(1)}$ :

$$f_{(1)}(x) = n(1 - F(x))^{n-1} f(x)$$

- Для **максимума**

$$F_{(n)}(x) = n \cdot F(x)^{n-1} \cdot f(x)$$

Рассмотрим следующий вопрос: Что если объем выборки ( $n$ ) достаточно велик? Возможно ли рассматривать распределение порядковой статистики как нормальное?

**Ответ:** Да, можно! Давайте посмотрим

### Вспомним, что означает формула

Функция распределения  **$k$ -й порядковой статистики**  $X_{(k)}$  выражается как:

$$F_{(k)}(x) = P(X_{(k)} \leq x) = \sum_{j=k}^n F(x)^j \cdot (1 - F(x))^{n-j} \cdot C_n^j$$

Это сумма биномиальных вероятностей — значит, **распределение  $X_{(k)}$  связано с биномиальным распределением.**

Введём обозначение:

$B_n$  = количество элементов выборки, не превышающих  $x$

Тогда  $B_n$  имеет биномиальное распределение:

$$B_n \sim \text{Bin}(n, F(x))$$

А порядковая статистика  $X_{(k)}$  — это **такое значение  $x$ , при котором  $B_n \geq k$ .**

ЦПТ утверждает, что если  $B_n \sim \text{Bin}(n, F(x))$ , то при большом  $n$  биномиальное распределение приближается к нормальному:

$$B_n \approx N(\mu = nF(x), \sigma^2 = nF(x)(1 - F(x)))$$

Тогда для  $X_{(k)}$ , используя аппроксимацию биномиала нормальным распределением, можно записать:

$$X_{(k)} \approx F^{-1}\left(\frac{k}{n} + N(0, \sigma^2)\right)$$
$$\sigma^2 = \frac{k}{n}\left(1 - \frac{k}{n}\right)$$

То есть, **если  $n$  достаточно велико, то  $X_{(k)}$  приближается к нормальному распределению с центром в квантиле  $F^{-1}\left(\frac{k}{n}\right)$ .**

### Что это означает на практике?

- Если  $n$  большое, то порядковая статистика  $X_{(k)}$  ведёт себя примерно как нормальная.
- Это полезно в задачах, связанных с экстремальными значениями, например, при анализе максимальных убытков в финансах или надежности оборудования.
- Однако асимптотическое нормальное распределение будет хорошим приближением только в середине выборки (для медианы или квантилей).

### 💡 Важно:

- Для крайних порядковых статистик (минимума и максимума) нормальное приближение хуже работает, потому что там мы приближаемся к распределениям экстремальных значений (Gumbel, Fréchet, Weibull).

### Итог

- Если  $n$  большое, можно рассматривать распределение порядковой статистики  $X_{(k)}$  как асимптотически нормальное.
- Центрируется оно в квантиле  $F^{-1}\left(\frac{k}{n}\right)$ .
- Но если рассматриваются крайние порядковые статистики (минимум или максимум), то лучше использовать теорию экстремальных значений, а не ЦПТ.

Реальное применение.

### Анализ экстремальных значений (Максимумы и минимумы)

Используется в страховании, финансах, инженерии и метеорологии.

### 📌 Страхование и анализ рисков

- **Максимальные убытки:** в страховой математике оценивают вероятность наступления редких, но очень дорогих событий (например, стихийных бедствий, аварий).
- **Value at Risk (VaR):** в финансах рассчитывают, какой убыток может произойти с заданной вероятностью (например, 5%-квантиль потерь).

### 💡 Пример:

Если банк хочет узнать, какой может быть **наибольший убыток** за месяц с вероятностью 99%, он рассматривает **99%-квантиль** распределения прибыли — а это **порядковая статистика!**

## Инженерия и надёжность

- **Минимальное время до отказа:** если у нас есть выборка из  $n$  устройств, мы анализируем  $X_{(1)}$  — время, когда **первое устройство сломается**.
- **Тестирование на прочность:** допустим, испытывают 100 бетонных балок. Инженеры интересуются **наименьшей прочностью** (потому что слабая балка может вызвать катастрофу).

### 💡 Пример:

Автопроизводитель испытывает 100 двигателей на прочность. Чтобы избежать массовых поломок, ему важно знать **минимальный срок службы**  $X_{(1)}$  — ведь именно он определяет гарантийный срок.

## Оценка квантилей и медианы

Порядковые статистики помогают находить **квантили** — значения, которые делят распределение на части.

### 📌 Медиана и квантильная оценка

- **Медиана** — устойчивая альтернатива среднему, так как не чувствительна к выбросам. В статистике часто рассматривают  $X_{(\frac{n}{2})}$  как **оценку медианы**.
- **Квантили** помогают в медицине и экономике (например, **90%-квантиль доходов** показывает, сколько зарабатывают **самые богатые 10%** людей).

### 💡 Пример:

Допустим, врач хочет узнать, какое **время восстановления** после операции типично для пациентов. **Среднее значение** может быть искажено (если у кого-то восстановление длилось 2 года). Вместо этого лучше взять **медиану**  $X_{(\frac{n}{2})}$ .

## Экстремальная метеорология

Порядковые статистики применяются в анализе **экстремальных погодных явлений**:

- **Максимальные температуры**  $X_{(n)}$  → при прогнозировании жары.
- **Минимальные температуры**  $X_{(1)}$  → при оценке рисков заморозков.
- **Наибольший уровень осадков** → для расчета наводнений.

### 💡 Пример:

Метеорологи изучают **наибольшие зафиксированные осадки за 50 лет**, чтобы оценить вероятность новых **аномальных дождей и наводнений**.

**Вывод**

Порядковые статистики используются в самых разных сферах:

Область	Что анализируют?	Как применяют?
Финансы	Максимальные по- тери	Оценка риска (VaR)
Страхование	Крупнейшие убытки	Расчёт страховых вы- плат
Инженерия	Минимальное время до отказа	Надёжность оборудова- ния
Медицина	Медианы времени восстановления	Оценка эффективности лечения
Метеорология	Экстремальные тем- пературы	Прогноз катастроф
Машинное обу- чение	Квантили признаков	Нормализация данных