



GeekBrains

Теория вероятностей и математическая статистика

Вебинары

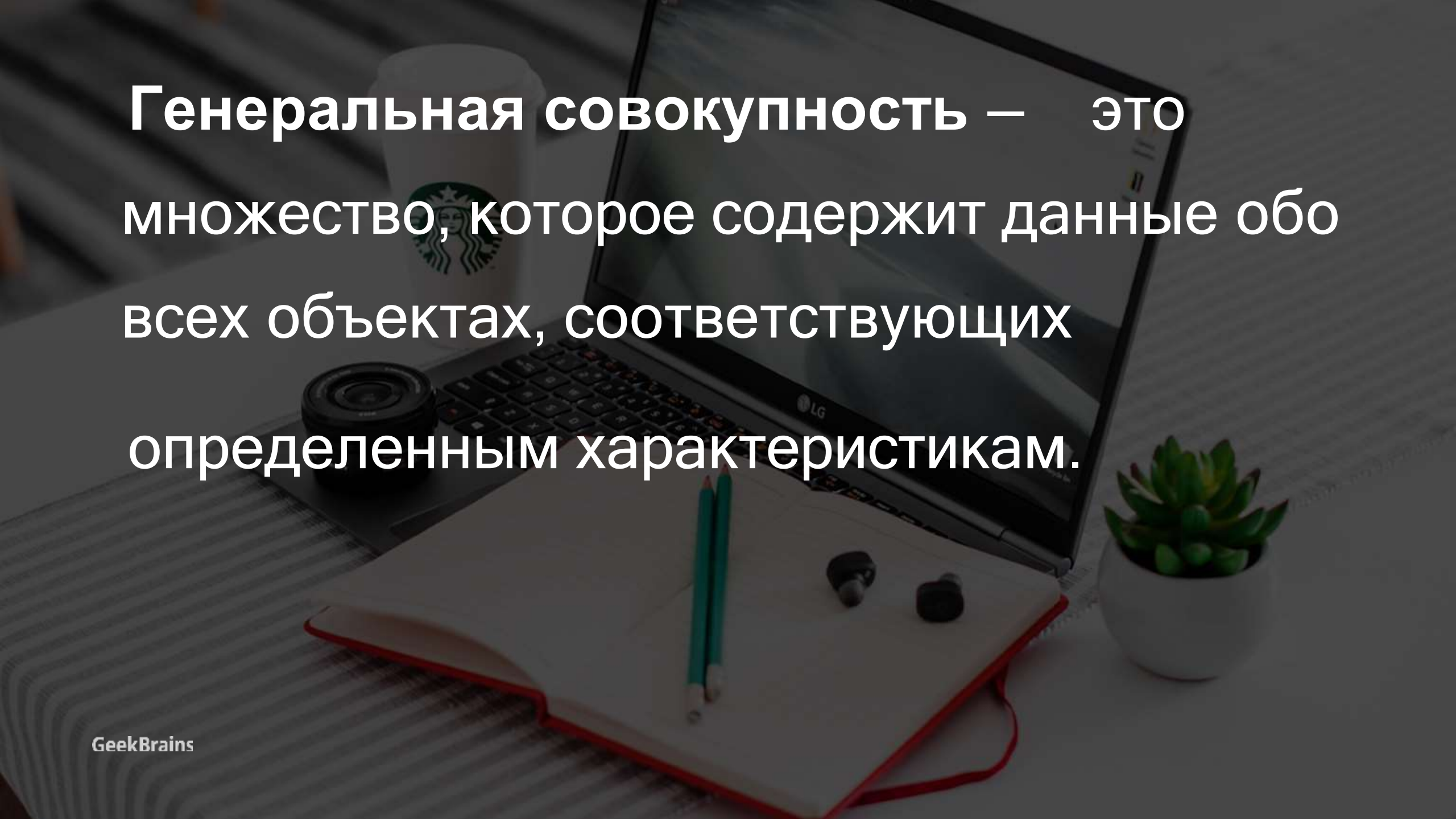


GeekBrains

Урок 3

Теория вероятностей и математическая статистика

Описательная статистика. Качественные и количественные
характеристики популяции. Графическое представление данных

A background image of a desk setup. It includes a black LG laptop, a white Starbucks cup, a camera lens, a red notebook with two green pencils, and a small potted succulent.

Генеральная совокупность — это множество, которое содержит данные обо всех объектах, соответствующих определенным характеристикам.

**Выборка — это случайным образом
выбранная часть генеральной
совокупности.**

Статистики

Для точечного оценивания параметров СВ используют различные статистики

Статистика – любая функция от выборки

Точечная оценка – наилучшее приближение к оцениваемому параметру

1. Среднее арифметическое / выборочное среднее

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

Является оценкой для математического ожидания

2. Выборочная дисперсия (смещенная оценка) - оценивает дисперсию случайной величины

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

оценка является несмещенной, если $M(Q^*) = Q$, где Q - это оцениваемый параметр

$$\sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$$

Число степеней свободы - число независимых переменных за вычетом числа промежуточных оценок, используемых в процессе построение оценки

$$\overline{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{X})^2$$

3. Среднеквадратическое отклонение (СКО)

$$\sigma_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{X})^2}$$

4. Мода — наиболее часто встречающееся в выборке значение.

5. Медиана — такое значение t , что половина элементов из выборки меньше, либо равна t , и, соответственно, половина больше, либо равна t .

Представляет собой середину в отсортированной выборке

6. Квантиль

Квантиль порядка α - такое число t_α , что « α процентов» всех элементов выборки меньше t_α , и соответственно, « $(1 - \alpha)$ процентов» элементов - больше t_α

Медиана = квантиль порядка 0,5

- Первый квартиль - квантиль порядка 0.25 (значение, которое не превышают 25% элементов выборки)
- Второй квартиль - квантиль порядка 0,5
- Третий квартиль - квантиль порядка 0,75

Дециль - квантиль порядка 0,1; 0,2 и тд

Перцентиль - аналог квантиля, но с использованием не доли, а процента (третий квартиль = 75% перцентиль)

Интерквартильное расстояние / размах — отрезок, равный разности третьей и первой квартили. Отрезок, в который попадают 50% выборки

Используется для измерения разброса значений выборки вокруг среднего

Квантиль случайной величины

Квантилем порядка α случайной величины X называют такое значение t_α , что

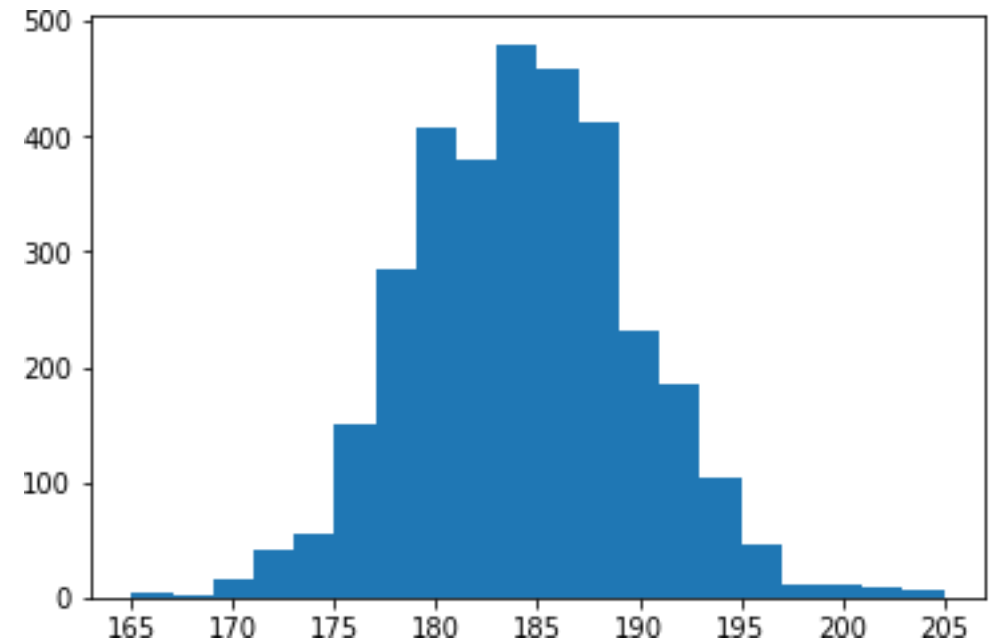
$$P(X \leq t_\alpha) = \alpha, \quad P(X \geq t_\alpha) = 1 - \alpha$$

В доле α всех случаев значение случайной величины X окажется меньше t_α , и соответственно, в доле $(1 - \alpha)$ случаев - больше t_α

Графическое представление данных

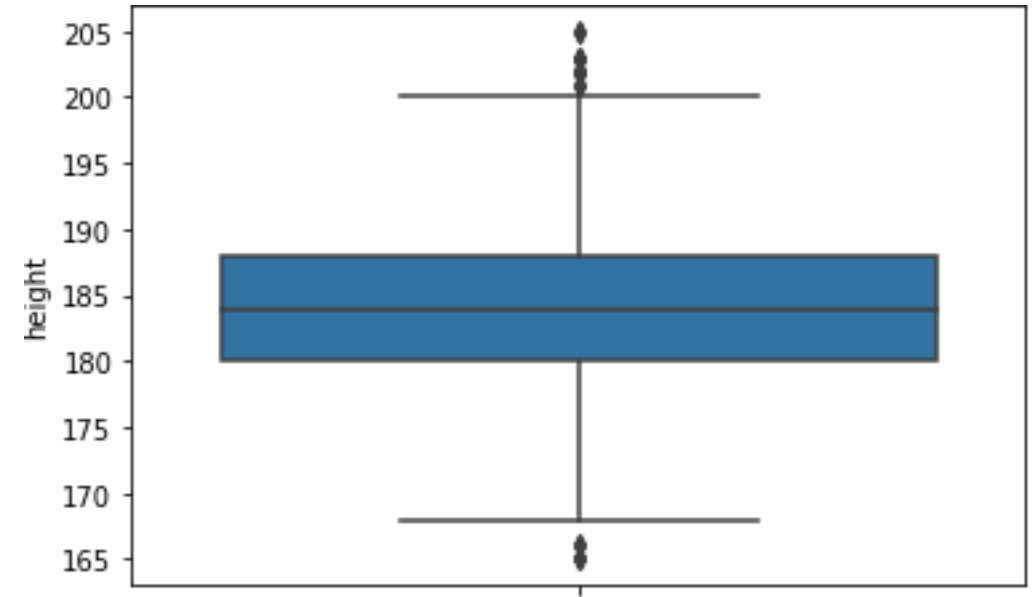
Для визуализации распределения значений выборки часто используется гистограмма

- По оси x откладываются все возможные значения из выборки.
- Вся ось разбивается на какое-то заданное число одинаковых отрезков.
- Для каждого отрезка вычисляется число значений выборки, которые лежат в этом



Boxplot / ящик с усами

- В самом ящике отмечены квартили Q1, Q2 (медиана), Q3.
- «Усы» здесь — границы отрезка $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$, где IQR — интерквартильное расстояние.
- Всё, что выходит за границы этого отрезка, считается выбросами (отмечены кружками).



ИТОГИ

1. Генеральная совокупность и выборка.
2. Оценка математического ожидания.
3. Дисперсия, среднее квадратичное отклонение. Смещенная и несмещенная оценка дисперсии.
4. Мода, медиана, квартиль, перцентиль, дециль, квантиль.
5. Графическое представление данных: гистограмма, boxplot.