



GeekBrains

# Теория вероятностей и математическая статистика

Вебинары





GeekBrains

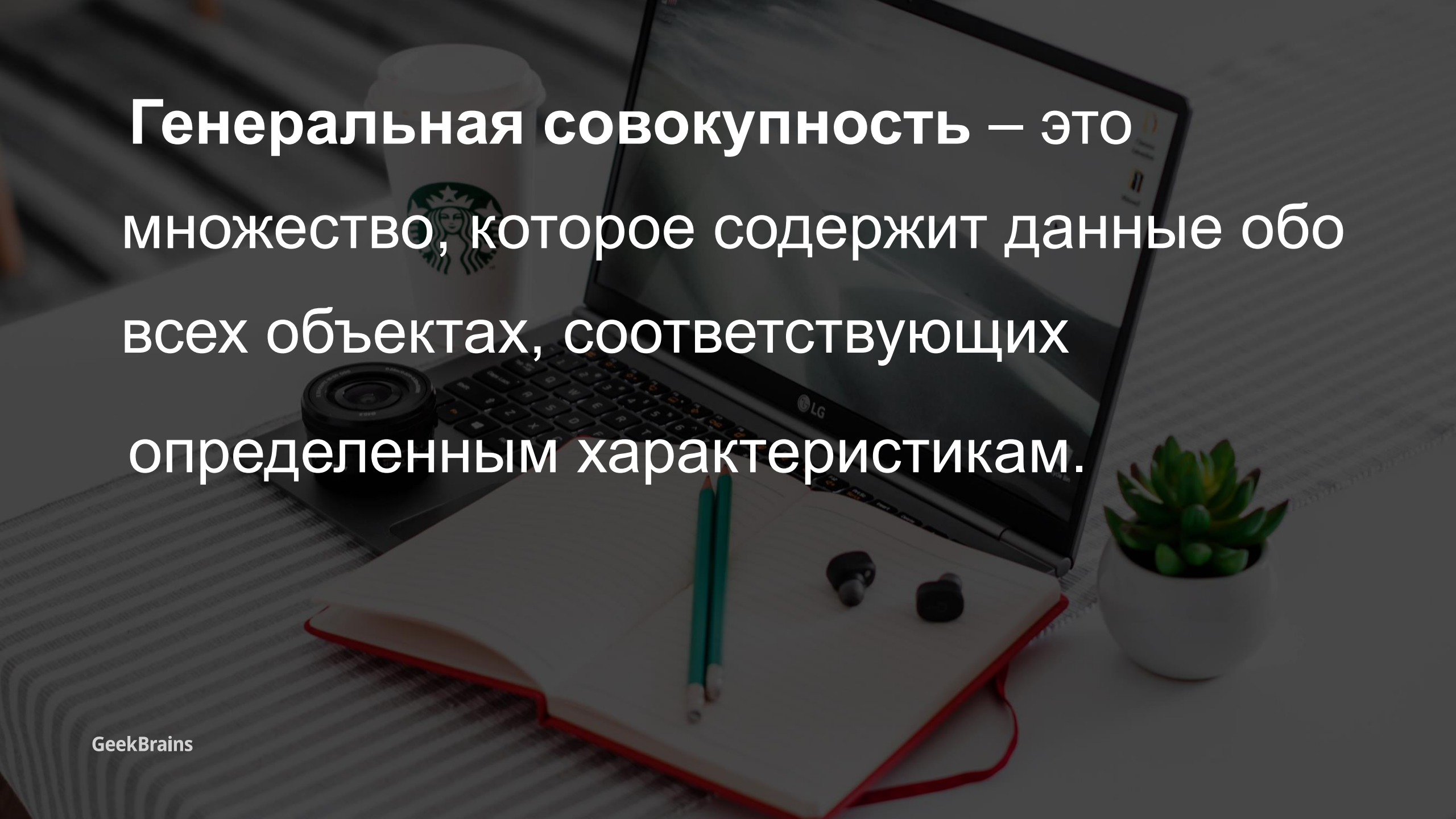
Урок 3

# Теория вероятности и математическая статистика

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

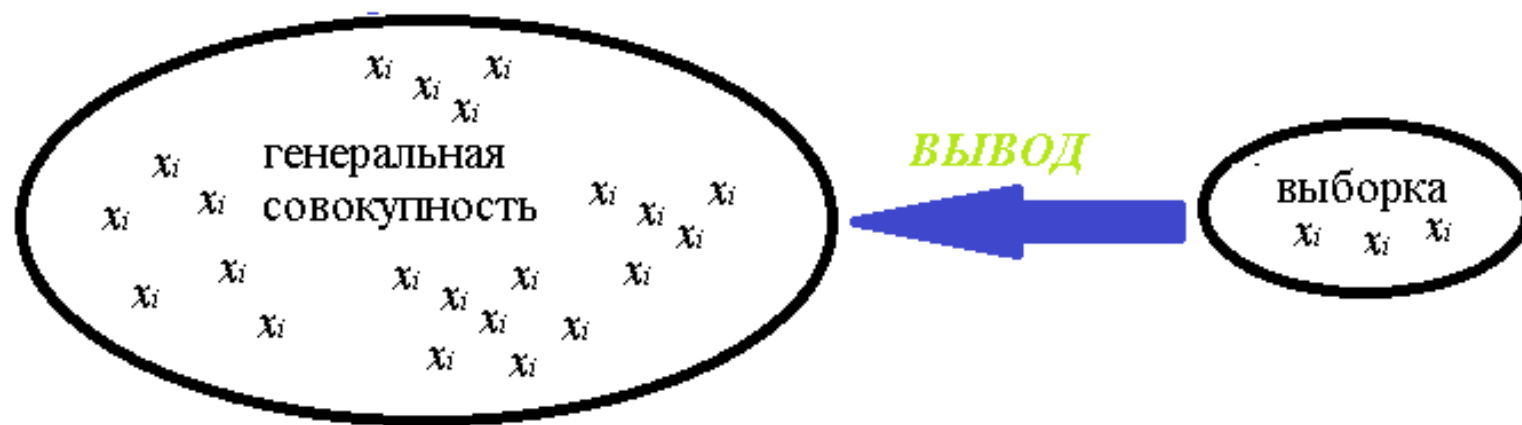
# На этом уроке мы изучим:

1. Генеральная совокупность и выборка
2. Математическое ожидание
3. Дисперсия, среднее квадратичное отклонение. Смещенная и несмещенная оценка дисперсии
4. Мода, медиана, квартиль, перцентиль, дециль, квантиль
5. Гистограмма, boxplot.

A background image of a desk setup. It includes a black LG laptop, a white Starbucks cup, a camera lens, an open notebook with two green pencils, and a small potted succulent. The text is overlaid on this image.

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

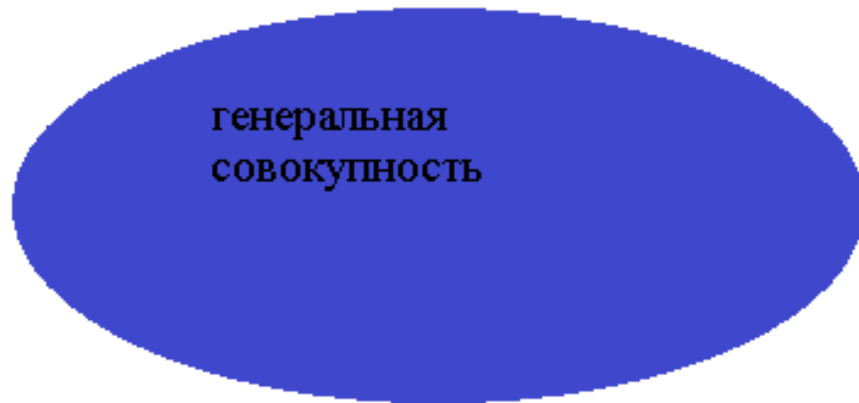
**Выборка - это случайным образом  
выбранная часть генеральной  
совокупности.**



$X$ -случайная  
переменная

$x_i$  - значения случайной  
переменной

Одной из основных характеристик генеральной совокупности является **математическое ожидание.**



$M(X)$  - мат.ожидание



$\mu$  - оценка  
математического  
ожидания

$$M(X) = \frac{1}{n} \sum_{i=1}^n x_i$$



**Математическое ожидание** — среднее значение случайной величины (распределение вероятностей стационарной случайной величины) при стремлении количества выборок или количества измерений (иногда говорят — количества испытаний) к бесконечности.

**Среднее арифметическое** одномерной случайной величины **конечного** числа испытаний обычно называют оценкой математического ожидания.



$M(X)$  - мат.ожидание

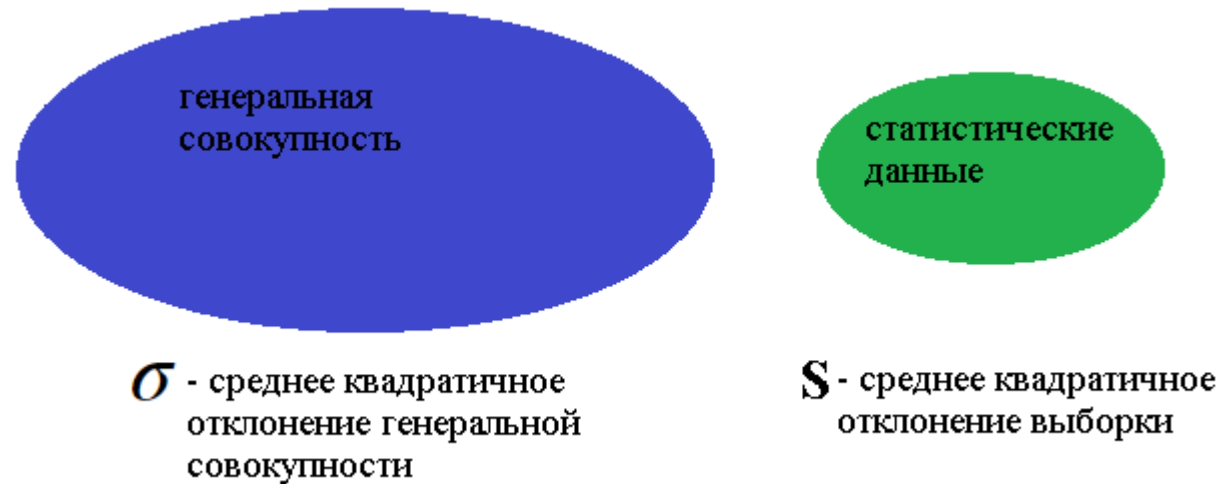


$\mu$  - оценка  
математического  
ожидания



**Среднее квадратичное отклонение** —еще важная характеристика.

Оно показывает, насколько далеко наблюдения могут быть "разбросаны" относительно среднего значения.



На практике обычно мы не можем рассчитать сигму, но мы можем это обойти, рассчитав среднее квадратичное отклонение выборки  $S$ .

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

**$m$**  - число объектов в генеральной совокупности

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

**$n$**  - объем выборки



**$\sigma$**  - среднее квадратичное отклонение генеральной совокупности



**$s$**  - среднее квадратичное отклонение выборки

**Дисперсия** равна среднему квадратичному отклонению, возведенному в квадрат.

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

**смещенная оценка**

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

**Несмещенная оценка дисперсии**

**Стандартное  
отклонение**

```
: 1 np.std(x)
: 0.42766809560686203

: 1 np.std(x, ddof=1)
: 0.4508017549014448
```

# Параметры генеральной совокупности и оценки

генеральная  
совокупность

$\mu$

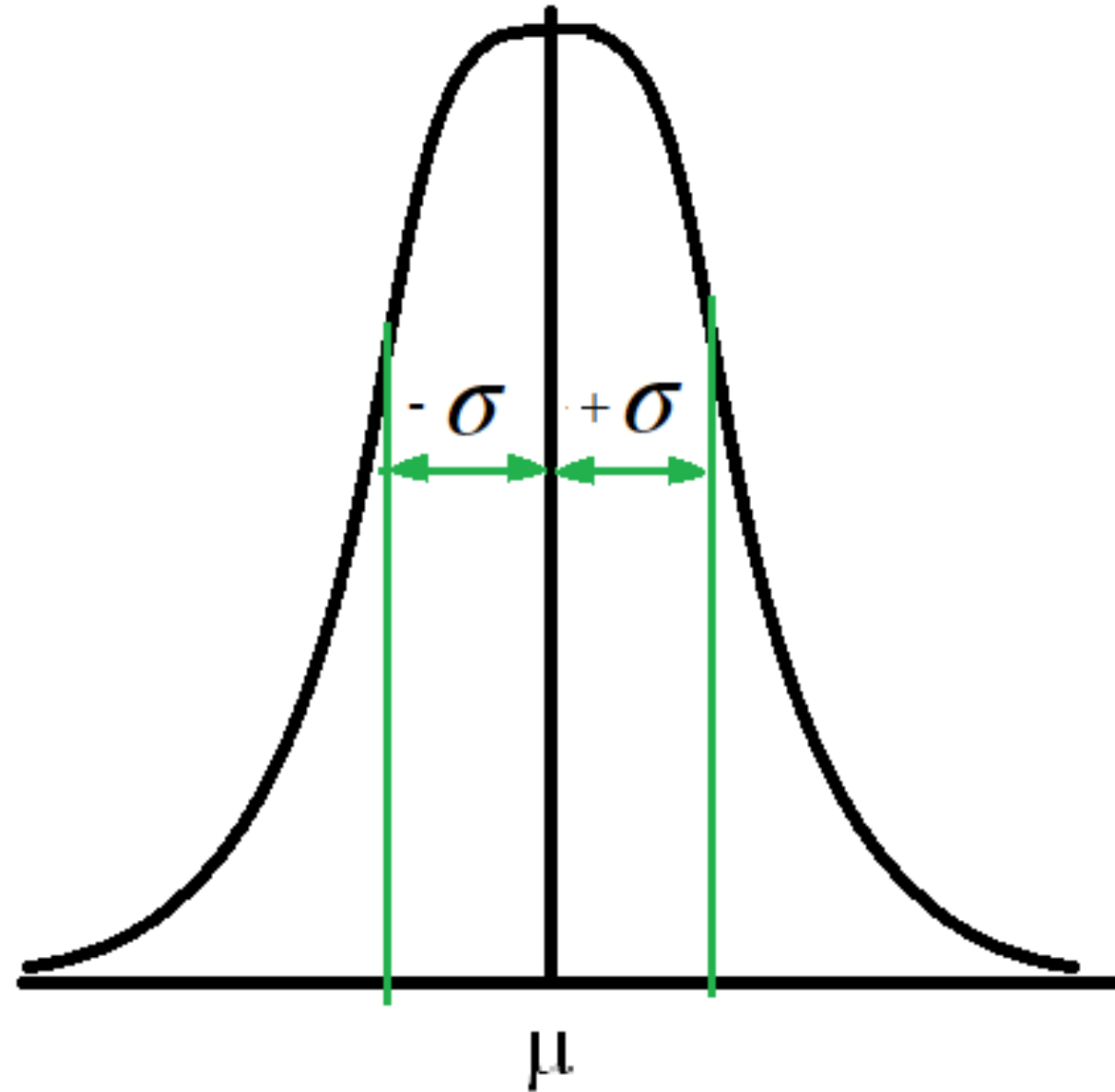
$\sigma$

статистические  
данные

$\mu$

$S$





**Медиана** - значение, которое делит выборку на две части, так что значения, которые меньше медианы, составляют половину (50%) выборки.

Нечетное число элементов в выборке

1 1 2 2 2 2 2 3 3 4 5  
50%                      50%

Медиана = 2 (шестое значение)

$N[(n+1)/2]$ ,

**n**- нечетное число измерений

Четное число элементов в выборке

1 3 7 8      Медиана =  $(3+7)/2 = 5$        $\frac{N[n/2] + N[n/2+1]}{2}$       **n**-четное число измерений

$(n[\text{len}(n)/2] + n[\text{len}(n)/2+1])/2$

```
In [31]: import numpy as np
```

```
In [46]: z=np.array([100,80,75,77,89,33,45,25,65,17,30,24,57,55,70,75,65,84,90,150])  
z
```

```
Out[46]: array([100,  80,  75,  77,  89,  33,  45,  25,  65,  17,  30,  24,  57,  
                55,  70,  75,  65,  84,  90, 150])
```

```
In [47]: z.shape
```

```
Out[47]: (20,)
```

```
In [48]: z.sort()
```

```
In [49]: z
```

```
Out[49]: array([ 17,  24,  25,  30,  33,  45,  55,  57,  65,  65,  70,  75,  75,  
                77,  80,  84,  89,  90, 100, 150])
```

```
In [50]: (z[9]+z[10])/2
```

```
Out[50]: 67.5
```

**Мода** - наиболее часто встречающееся в выборке значение.

	буквы									
частота	a	b	c	d	e	f	g	h	i	j
	1	2	2	2	5	3	3	2	2	1



**Первый квартиль** - такое значение, что 25% наблюдений в выборке не превышают эту величину.

**Второй квартиль** - синоним медианы.

**Третий квартиль** - такое значение, что 75% наблюдений в выборке не превышают эту величину.

**Интерквартильное расстояние** - отрезок, равный разности 3-го и 1-го квартиля.

```
In [15]: z= np.array([1,2,4,2,1,5,7,2,3,5,7,8,9])  
z
```

```
Out[15]: array([1, 2, 4, 2, 1, 5, 7, 2, 3, 5, 7, 8, 9])
```

```
In [22]: n=len(z)  
n
```

```
Out[22]: 13
```

```
In [17]: z.sort()
```

```
In [18]: z
```

```
Out[18]: array([1, 1, 2, 2, 2, 3, 4, 5, 5, 7, 7, 8, 9])
```

```
In [19]: ## Если  $n*k/100$ -целое число,то  $k$ -я персентиль это среднее значений под номерами  $n*k/100$  и  $n*k/100+1$ 
```

```
In [21]: ## Если  $n*k/100$ -не целое число,то  $k$ -я персентиль совпадает с измерением  $j+1$ , где  $j$  -максимальное целое число  $< n*k/100$ 
```

```
In [23]: ## Посчитаем 25ю персентиль
```

```
In [26]: k=25
```

```
In [27]: n*k/100
```

```
Out[27]: 3.25
```

```
In [28]: j=3
```

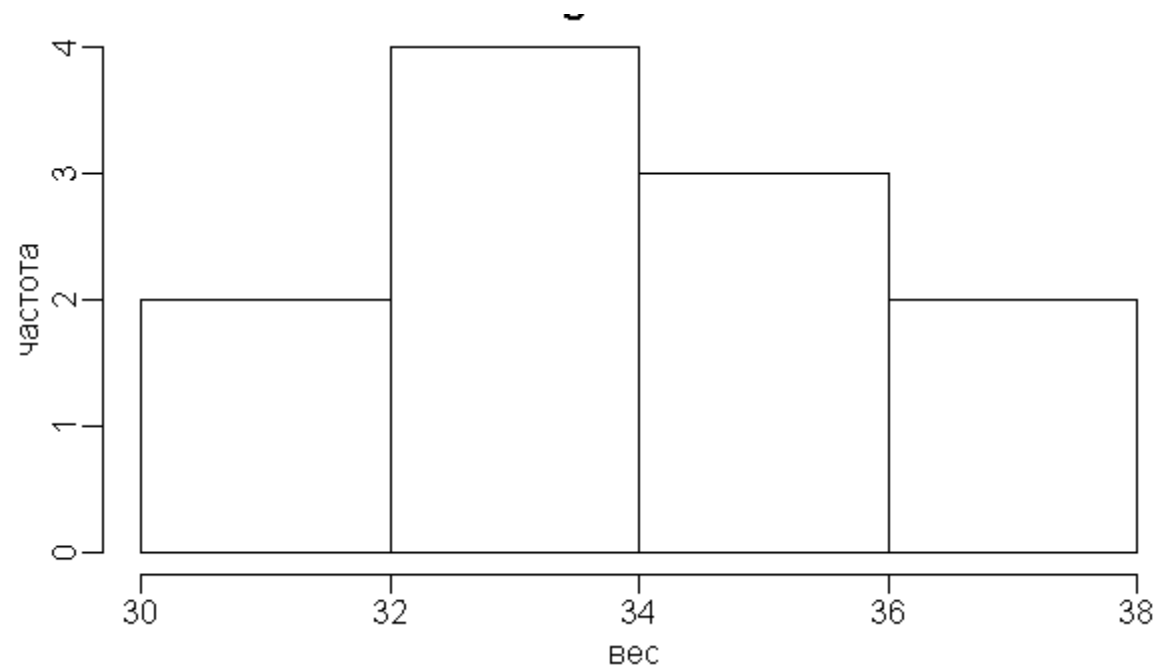
25 персентиль соответствует  $j+1=4$ , т.е. 4-му значению ,котре равно

```
In [35]: z[3]
```

```
Out[35]: 2
```

32.34566, 34.96313, 33.87, 35.61900, 35.60872, 33.11, 32.78, 30.11787, 30.45296, 36.41410, 37.86643

$$R = X_{\max} - X_{\min} = 37.86643 - 30.45296$$



Помимо **квартилей**, в статистике используются:

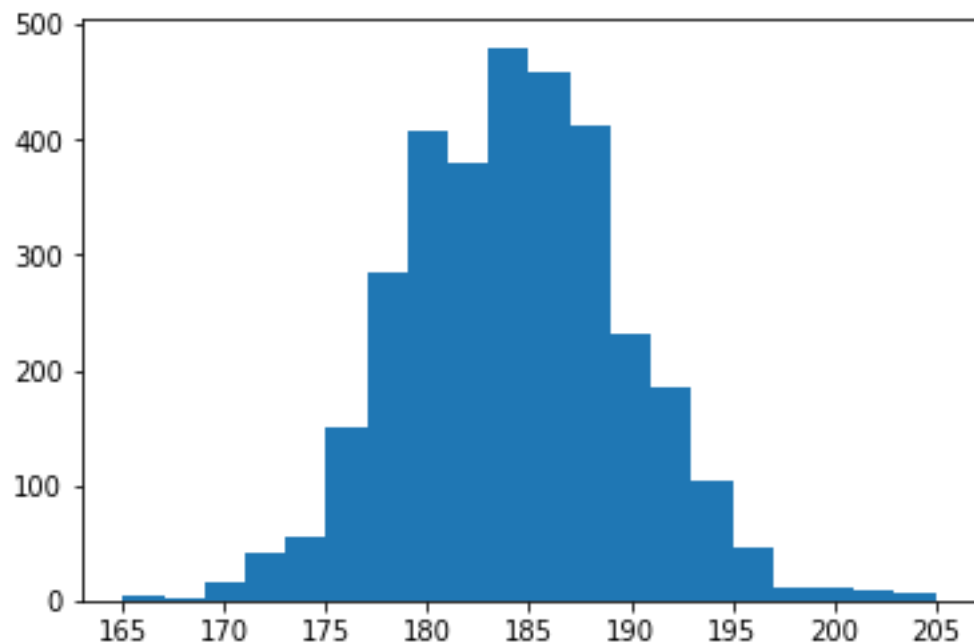
**-перцентили,**

**-децили**

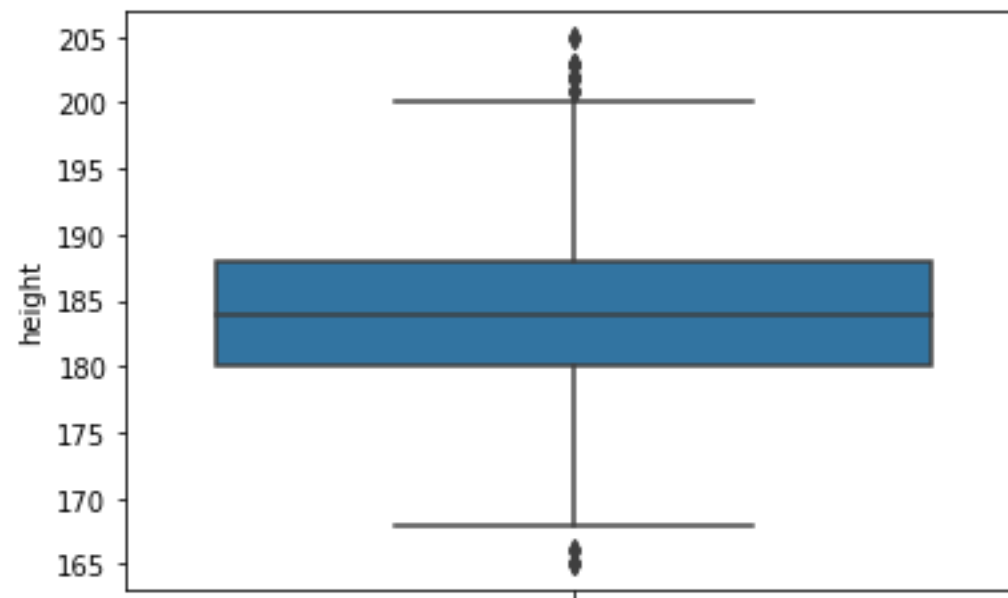


# Графическое представление данных

Гистограмма



Boxplot



$$X1 = Q1 - 1.5 * (Q3 - Q1) ; X2 = Q3 + 1.5 * (Q3 - Q1)$$

# Итоги

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