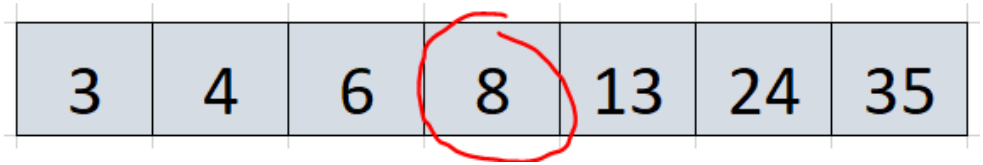
Метрики центрального положения

Среднее (mean)
Сумма всех значений, деленная на количество значений или среднее арифметическое.

Медиана (median)
Середина в отсортированных данных.



Мода (mode)
Значение, которое встречается наиболее часто.

мода=3

Мода часто употребляется для текстовых данных. Например: цвета автомобилей — белый, чёрный, синий металлик, белый, синий металлик, белый. Какая мода?

Выброс (outlier)
Значение данных, которое сильно отличается от большинства данных.

Метрики оценки вариативности

Размах (range) Разница между самым большим и самым малым значениями в наборе данных.

Математи́ческое ожида́ние — среднее (взвешенное по вероятностям возможных значений) значение случайной величины. На практике математическое ожидание обычно оценивается как среднее арифметическое наблюдаемых значений случайной величины (выборочное среднее, среднее по выборке).

Дисперсия (variance) средний квадрат отклонений индивидуальных значений признака от их средней величины. Еще называют: *среднеквадратическое отклонение, среднеквадратическая ошибка*.

Стандартное отклонение (standard deviation) Квадратный корень из дисперсии. И в отличие от дисперсии оно показывает реальное среднее значение наших отклонений от среднего значения по выборке.

Процентиль – например, **75**-й процентиль – это число, ниже которого находится **75**% всех наблюдений.

Квартили - такие точки, которые делят наше распределение на 4 равные части (25%, 50% и 75%)

Квантили *- такие значения признака, которые делят упорядоченные данные на некоторое число равных частей.*

	Age
count	714.000000
mean	29.699118
std	14.526497
min	0.420000
25%	20.125000
50%	28.000000
75%	38.000000
max	80.000000

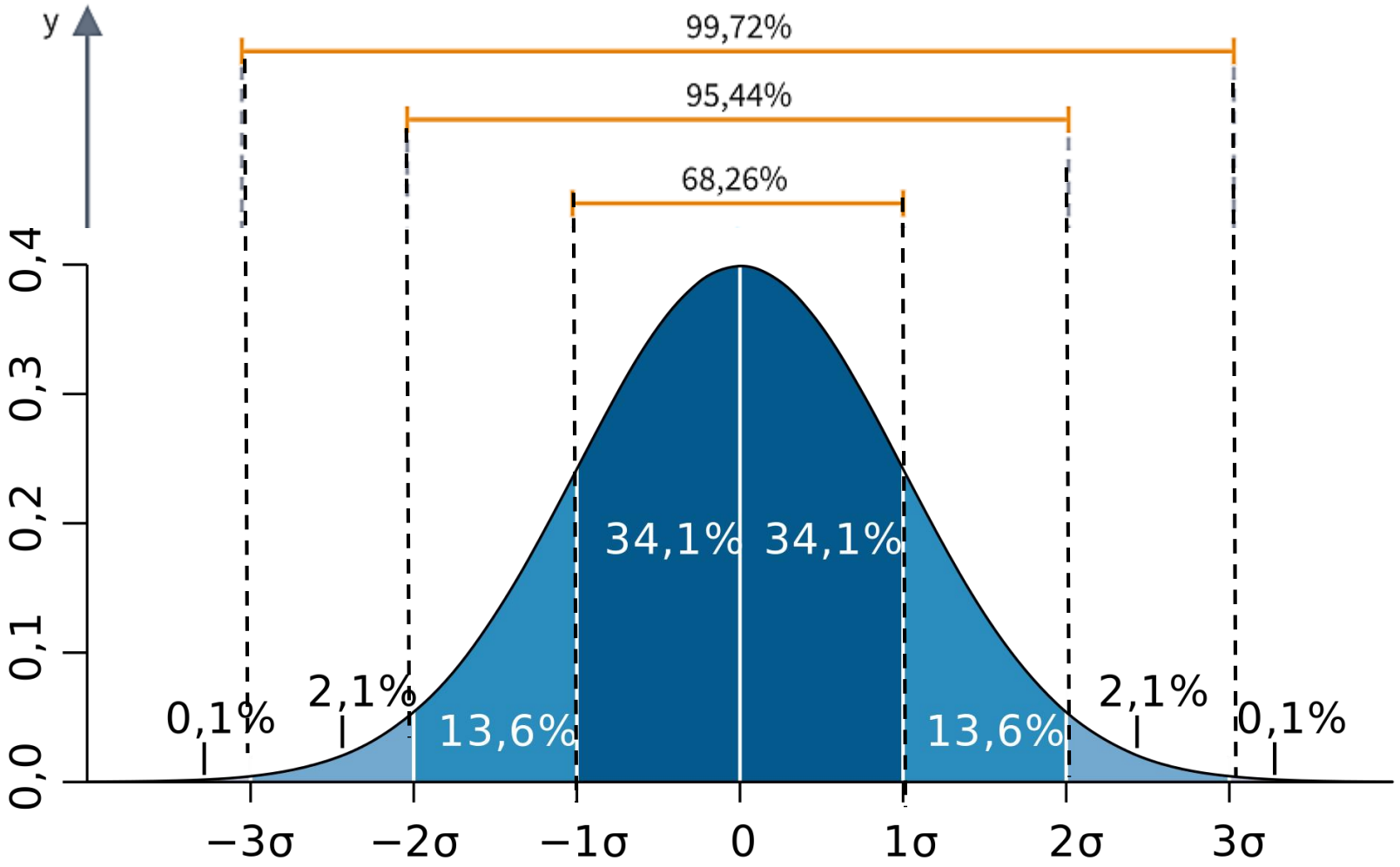
Процентиль на примере

Допустим, у нас есть массив возрастов пассажиров на Титанике.

Из примера 25-й **процентиль** равен 20, это означает, что 25% людей моложе 20 лет.

Нормальное распределение

1. Среднее, медиана и мода равны.
2. Кривая симметрична
3. Площадь под кривой равна 1.

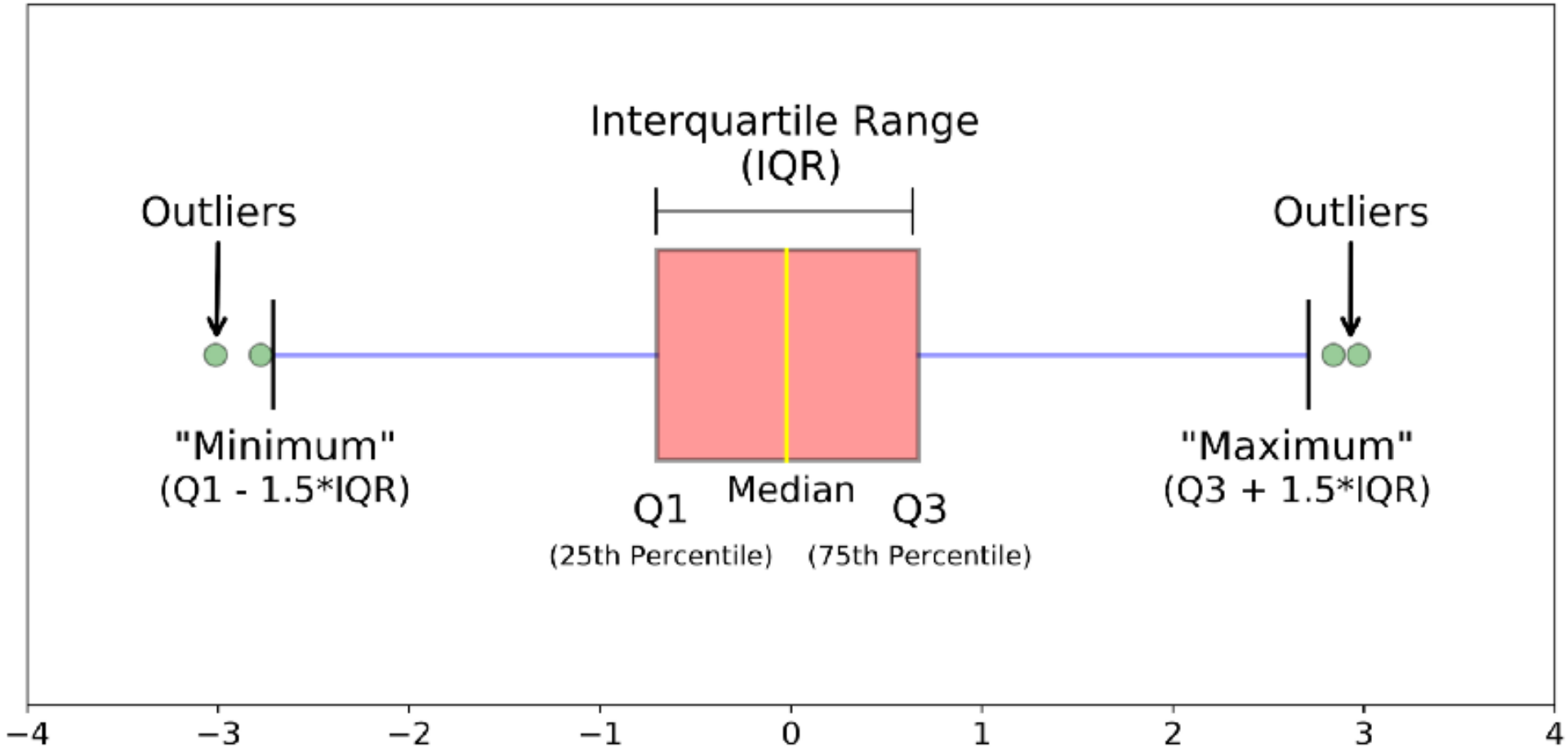


Правило трёх сигм:

вероятность того, что случайная величина примет значение, отклоняющееся от математического ожидания больше чем на три среднеквадратических отклонения, не превышает 0,28%, т. е. пренебрежимо мала.

δ — среднеквадратичное отклонение

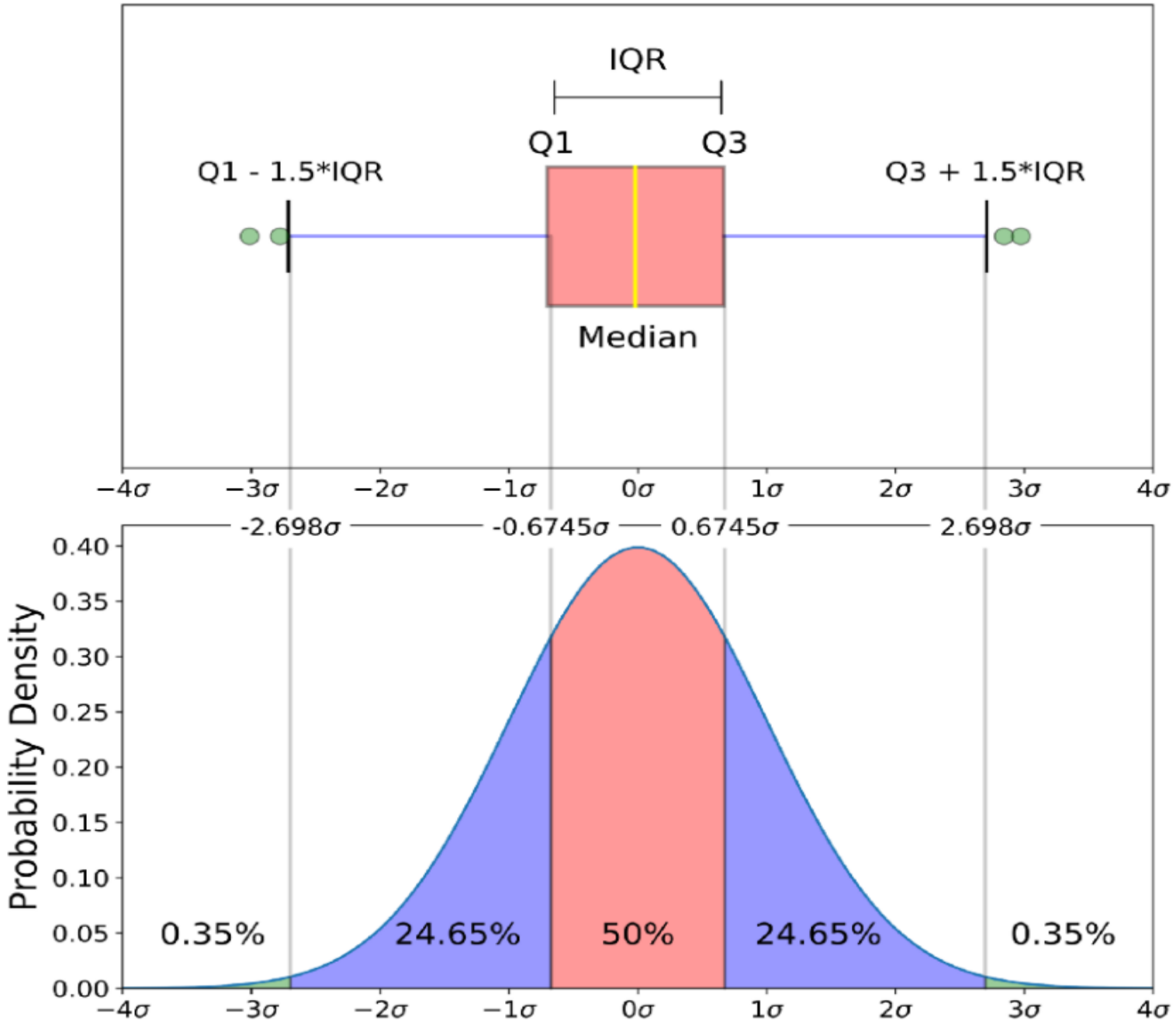
box plot Ящик с усами Диаграмма размаха



IQR - размах

Q1 – 25 перцентиль (1й квартиль)

Q3 – 75 перцентиль (3й квартиль)



Box-plot это еще одна визуализация распределения случайной величины

На рисунке можно наглядно увидеть сколько данных содержится в ящике, сколько в усах и сколько за пределами усов.