



ОСНОВЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ для будущего врача

Размах • Среднее • Ошибка среднего • Медиана • Квартили

На примере инфекционных болезней

Шип С.А. (для личного пользования)

1

Раздел 1

Размах значений

Шип С.А. (для личного пользования)

Размах: формула

Размах = максимум – минимум = Max - Min

Самая простая мера разброса значений

Показывает, насколько сильно данные
«разъезжаются»

Пример: инкубационный период — от сколько до
сколько дней заболели пациенты

**Подводный камень: размах определяет одно крайнее
значение**

Размах: пример — инкубационный период COVID-19

Данные, дней: 2, 3, 4, 5, 5, 6, 7, 8, 14 ($n = 9$)

Максимум (Max) = 14, минимум (Min) = 2

Размах = 14 – 2 = 12 дней

Вывод: от заражения до первых симптомов прошло от 2 до 14 дней

14 дней — редкое (но возможное!) значение: оно и задало весь размах

Задание 1

Задача. Инкубационный период ветряной оспы (дни):

10, 12, 13, 14, 14, 15, 16, 17, 21

Найдите размах.

Подсказка: сначала найдите максимум и минимум.

Ответ — на следующем слайде

Ответ к заданию 1

Максимум (Max) = 21, минимум (Min) = 10

Размах = 21 – 10 = 11 дней

Комментарий: 21 день — редкое значение (возможно у ослабленных пациентов).

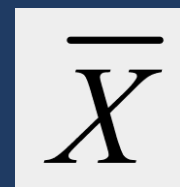
Выбросы нельзя просто вычеркнуть: в инфекционной практике крайние сроки бывают реальны.

2

Раздел 2

Среднее значений
(среднее
арифметическое
значение)

Mean (M)

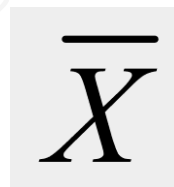


Шип С.А. (для личного пользования)

Среднее (среднее арифметическое значение) (Mean): формула

Среднее = сумма всех значений / число значений

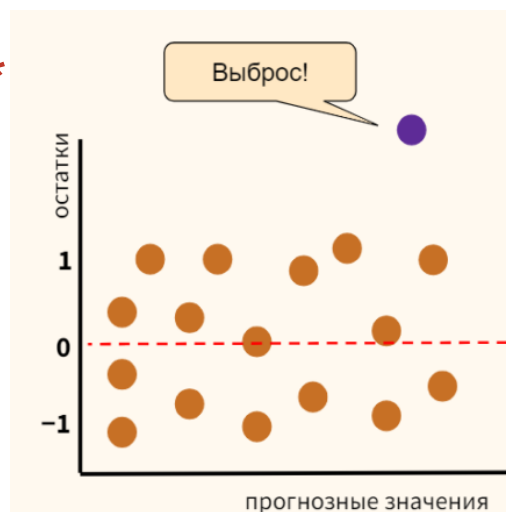
Обозначение — «x с чертой» (x-бар)



Отвечает на вопрос: «сколько в среднем» — случаев в день, дней лихорадки и т. д.

Подводный камень:
среднее чувствительно к выбросам*

* - **Выбросы в статистике** — это значения в наборе данных, которые заметно отличаются от остальных наблюдений, резко выделяясь на общем фоне. Проще говоря, это точки, которые не вписываются в общую картину распределения.



Среднее: пример — грипп

Новые случаи гриппа за неделю, в день:

10, 12, 14, 12, 16, 14, 13 (n = 7)

Сумма = 91

**Среднее (Mean) = $91 / 7 = 13$
случаев в день**

Вывод: за неделю эпидемия даёт в среднем 13 новых случаев ежедневно

Задание 2

Задача. Госпитализации с ротавирусной инфекцией за 6 дней:

7, 8, 10, 10, 12, 7

Найдите среднее число (Mean) госпитализаций в день.

Ответ — на следующем слайде

Ответ к заданию 2

$$\text{Сумма} = 7 + 8 + 10 + 10 + 12 + 7 = 54$$

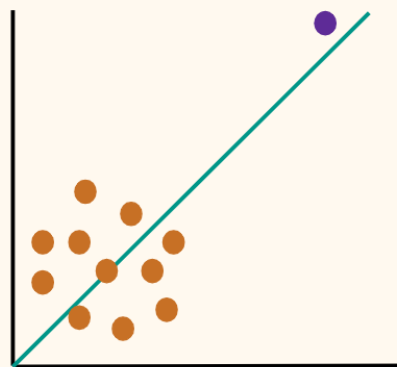
$$\text{Среднее (Mean)} = 54 / 6 = 9$$

госпитализаций в день

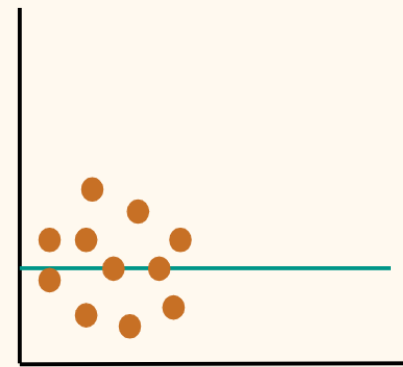
Комментарий: если бы один день дал 60 случаев, среднее подскочило бы до ~17 —

поэтому вместе со средним всегда смотрите на разброс и медиану.

** - **Выбросы в статистике** — это значения в наборе данных, которые заметно отличаются от остальных наблюдений, резко выделяясь на общем фоне. Проще говоря, это точки, которые не вписываются в общую картину распределения.*



с выбросом



без выброса

3.1

Раздел 3.1

Стандартная

ошибка среднего

Standard Error of the Mean

(SEM)

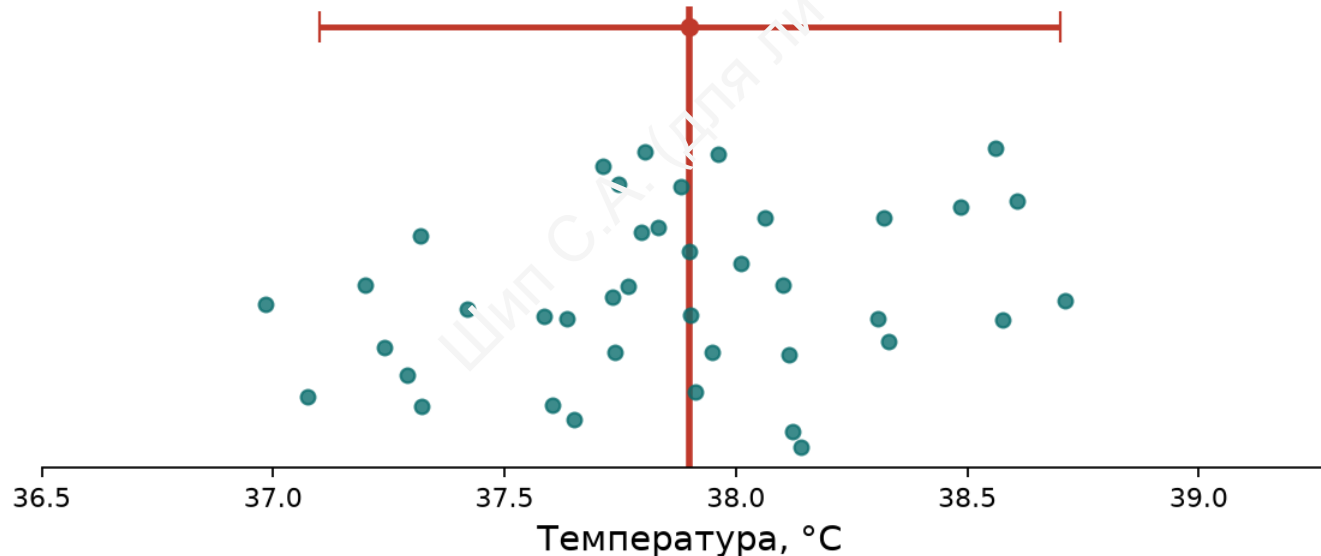
Шип С.А. (для личного пользования)

SEM: откуда берётся разброс

Каждая точка — среднее одной «виртуальной выборки» из 9 пациентов.

Точек много, а истина одна — её мы и ищем, уменьшая SEM.

Разные выборки дают разные средние — они «пляшут» вокруг истины



SEM (Standard Error of the Mean): **формула и смысл**

$$\text{SEM} = s / \sqrt{n}$$

s — стандартное отклонение (разброс отдельных пациентов)

n — число пациентов в выборке

SEM — насколько точно среднее выборки «угадывает» среднее популяции

Больше пациентов $\rightarrow$ меньше SEM $\rightarrow$ точнее оценка

SEM: пример — температура при гриппе

Пример:

Исследование: $n = 9$ пациентов, средняя температура = $37.9\text{ }^{\circ}\text{C}$, $s = 1.2\text{ }^{\circ}\text{C}$

$$\text{SEM} = 1.2 / \sqrt{9} = 1.2 / 3 = 0.4\text{ }^{\circ}\text{C}$$

Раздел 3.2

3.2

95% Доверительный Интервал

(95% ДИ)

Шип С.А. (для личного пользования)

«Грубый» Доверительный Интервал для SEM (95% ДИ): пример — температура при гриппе

*Пример: Исследование: $n = 9$ пациентов, средняя температура = $37.9\text{ }^{\circ}\text{C}$, $s = 1.2\text{ }^{\circ}\text{C}$. **$SEM = 1.2 / \sqrt{9} = 1.2 / 3 = 0.4\text{ }^{\circ}\text{C}$***

Грубое правило 95%: ДИ $\approx$ среднее $\pm 2 \times SEM$

ДИ (грубо) $\approx 37.9 \pm 2 \times 0.4 \rightarrow$ от 37.1 до $38.7\text{ }^{\circ}\text{C}$

Смысл: с вероятностью $\sim 95\%$ истинное среднее по популяции лежит в этом коридоре

95% ДИ – как считать?

95% ДИ (грубо)	среднее $\pm 2 \times SEM$	правило «большого пальца»
95% ДИ (точно)	среднее $\pm t \times SEM$	t — по таблице, $df = n - 1$

«Точный» 95% Доверительный Интервал для среднего (M): коэффициент t

Пример: Исследование: $n = 9$ пациентов, средняя температура = $37.9\text{ }^{\circ}\text{C}$, $s = 1.2\text{ }^{\circ}\text{C}$. **$SEM = 1.2 / \sqrt{9} = 1.2 / 3 = 0.4\text{ }^{\circ}\text{C}$**

95% ДИ = среднее $\pm t \times SEM$

t — табличный* коэффициент Стьюдента, зависит от $df = n - 1$

При $n = 9$: $df = 8$, $t = 2.306$ — больше, чем «2» из грубого правила

Почему: при малых выборках среднее «пляшет» сильнее $\rightarrow$ доверяемся осторожнее

При больших n (30+) $t \approx 1.96$ — правило « $\pm 2 SEM$ » снова работает

*

$df = n - 1$	t
4	2.776
8	2.306
15	2.131
24	2.064
30+	≈ 1.96

ДИ (точно): $37.9 \pm 2.306 \times 0.4 \rightarrow$ от 37.0 до $38.8\text{ }^{\circ}\text{C}$

Грубое правило давало (37.1; 38.7) — чуть уже, чем положено.

95% ДИ – как считать?

95% ДИ (грубо)	среднее $\pm 2 \times SEM$	правило «большого пальца»
95% ДИ (точно)	среднее $\pm t \times SEM$	t — по таблице, $df = n - 1$

Задание 3

Задача. В группе из $n = 16$ пациентов стандартное отклонение температуры $s = 1.5$ °C.

Найдите стандартную ошибку среднего.

Подсказка: $SEM = s / \sqrt{n}$. Чему равен $\sqrt{16}$?

Ответ — на следующем слайде

Ответ к заданию 3

$$\sqrt{16} = 4$$

$$\text{SEM} = 1.5 / 4 = 0.375 \approx 0.38 \text{ }^{\circ}\text{C}$$

Комментарий: 16 пациентов вместо 9 сократили ошибку вдвое —

SEM уменьшается как $\sqrt{n}$: чтобы снизить её вдвое, пациентов нужно в 4 раза больше.

Раздел 3.3

3.3

Стандартное отклонение

Шип С.А. (для личного пользования)

(S)

Стандартное отклонение: что это

s — «средняя» величина отклонения значений от среднего арифметического.

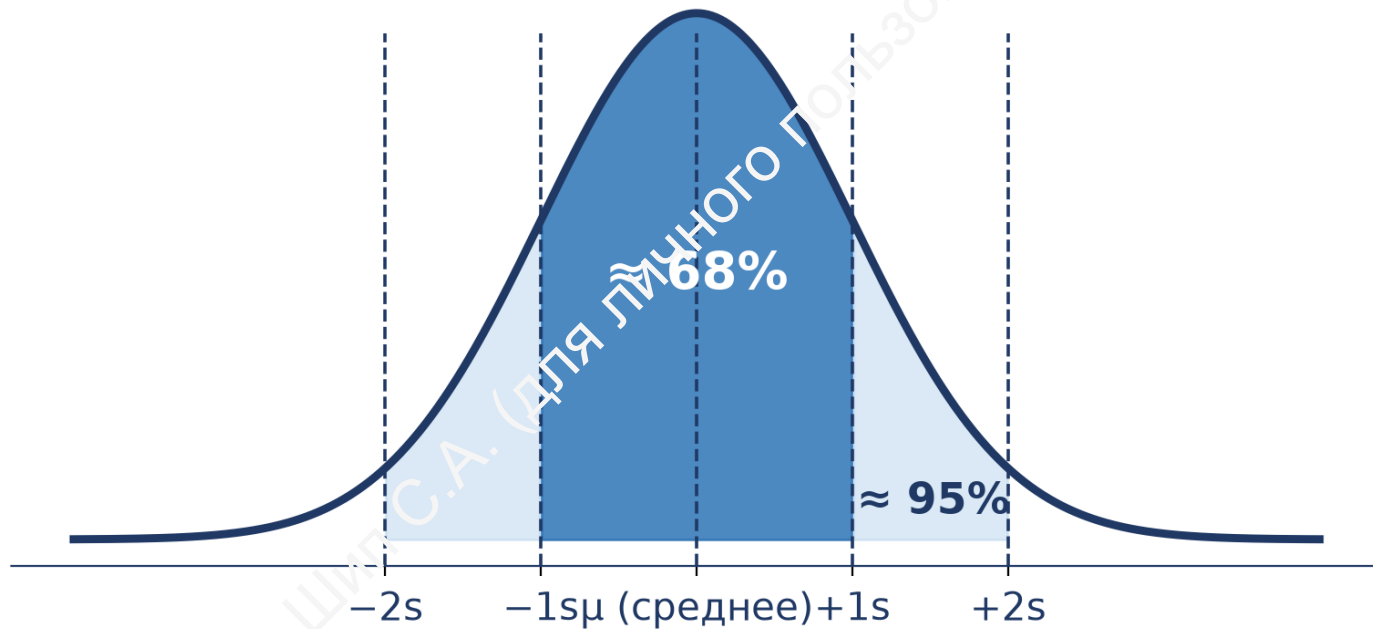
Формула:

$$s = \sqrt{ \sum (x_i - \bar{x})^2 / (n - 1) }$$

Почему $n - 1$, а не n : так оценка получается несмещённой, мы теряем одну «степень свободы», потому что уже использовали среднее. При больших n разницы почти нет.

s и правило двух сигм

Нормальное распределение: зоны $\pm 1s$ и $\pm 2s$



Примерно 68% значений лежат в пределах $\pm 1s$ от среднего, $\approx 95\%$ — в пределах $\pm 2s$.

Пример: длительность лихорадки при гриппе

У 5 пациентов лихорадка длилась: 4, 5, 5, 5, 6 дней.

Среднее $\bar{x} = 5$ дней; отклонения: $-1, 0, 0, 0, +1$; сумма квадратов = 2.

$$s^2 = 2 \div (5 - 1) = 0,5 \rightarrow s = \sqrt{0,5} \approx 0,7 \text{ дня.}$$

Читаем: значения в среднем отклоняются от 5 примерно на 0,7 дня. Маленький s — данные кучные.

Задание

ЗАДАНИЕ

Длительность госпитализации при гепатите А у 5 пациентов: 6, 7, 8, 9, 10 дней.

Вопрос: найдите среднее и стандартное отклонение s .

Подсказка: сначала среднее, затем отклонения от него и их квадраты.

Ответ — на следующем слайде.

Ответ к заданию

ОТВЕТ

Среднее = $(6+7+8+9+10) \div 5 = 8$ дней.

Отклонения: $-2, -1, 0, +1, +2$; сумма квадратов = $4 + 1 + 0 + 1 + 4 = 10$.

$$s^2 = 10 \div 4 = 2,5 \rightarrow s = \sqrt{2,5} \approx 1,6 \text{ дня.}$$

Комментарий: s здесь заметно больше, чем в примере с гриппом, — сроки госпитализации «размазаны» сильнее.

4

Раздел 4

Медиана значений

Median

(Me)

Шип С.А. (для личного пользования)

Медиана: определение

Медиана — значение в середине упорядоченного ряда

Шаги: 1) упорядочить данные 2) взять центральное значение

Если n чётное — медиана $\approx$ среднее двух центральных значений

Не боится выбросов — в отличие от среднего

Медиана: пример — длительность госпитализации

Дни в стационаре: 5, 6, 7, 7, 8, 8, 9, 10, 20 (n = 9, уже упорядочено)

Центральное (5-е) значение = 8

Медиана = 8 дней

А среднее = $80 / 9 \approx 8.9$ дней — его тянет вверх один «долгий» пациент (20 дней)

Врач видит типичного пациента — 8 дней, а не 8.9

Задание 4

Задача. Длительность лихорадки при гриппе (дни):

3, 4, 4, 5, 5, 6, 7

Найдите медиану.

Ответ — на следующем слайде

Ответ к заданию 4

Данные упорядочены, $n = 7$ — нечётное

Медиана = 4-е значение = 5 дней

Если бы n было чётным (например, 6 значений):

медиана = среднее двух центральных, например $(5 + 5)/2 = 5$

5

Раздел 5

Квартили (нижний и верхний)

Шип С.А. (для личного пользования)

Q1 – нижний квартиль

Q3 – верхний квартиль

Квартили: определение

Квартили режут упорядоченный ряд на 4 части по 25%

Q1 (нижний) — 25% значений ниже него

Q2 — это медиана (Me) (50%)

Q3 (верхний) — 75% значений ниже него

$IQR^* = Q3 - Q1$ — «внутренний» размах, свободный от крайних значений

*** - *Interquartile Range (IQR)*, или межквартильный размах**

Квартили: пример — сыпь и госпитализация

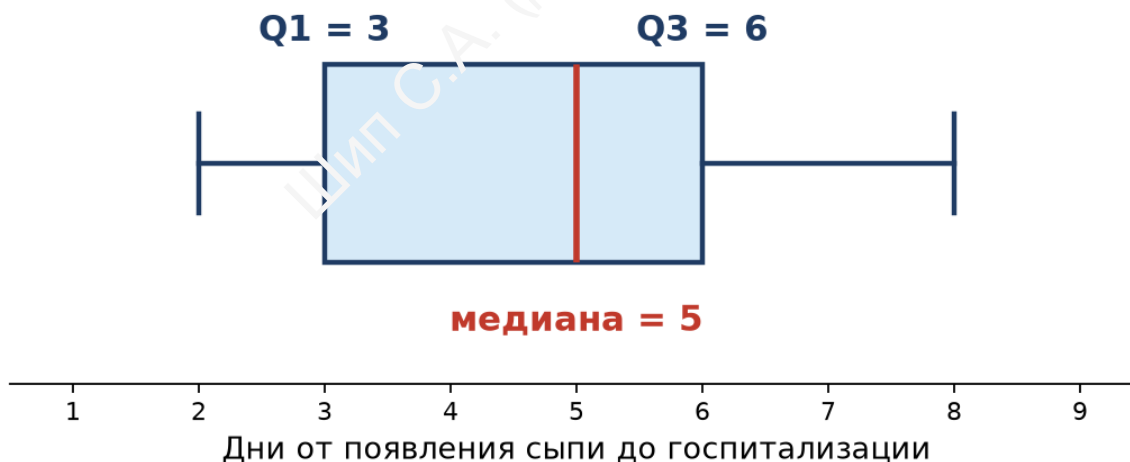
Дни от появления сыпи до госпитализации:

2, 3, 3, 4, 5, 5, 6, 7, 8 ($n = 9$, упорядочено)

$Q1 = 3$ -е значение = 3 дня • $Q3 = 7$ -е значение = 6 дней

Медиана ($Q2$) = 5-е значение = 5 дней

IQR



Задание 5

Задача. Утренняя температура 9 пациентов с ОРВИ (°C):

37.0, 37.1, 37.2, 37.4, 37.6, 37.8, 38.0, 38.3, 38.6

Найдите Q1, Q3 и IQR.

Подсказка: при $n = 9$ квартили — это 3-е и 7-е значения.

Ответ — на следующем слайде

Ответ к заданию 5

$$Q1 = 3\text{-е значение} = 37.2 \text{ }^{\circ}\text{C}$$

$$Q3 = 7\text{-е значение} = 38.0 \text{ }^{\circ}\text{C}$$

$$IQR = 38.0 - 37.2 = 0.8 \text{ }^{\circ}\text{C}$$

Вывод: средняя половина пациентов укладывается в узкий коридор $0.8 \text{ }^{\circ}\text{C}$

6

Раздел 6

Задания для самопроверки (с ответами)*

Шип С.А. (для личного пользования)

** Если вы решите всё верно — статистика покорена.
Если где-то ошиблись — вернитесь к нужному разделу.*

Задания для самопроверки*

Задача 1. Инкубационный период гепатита А (дни): 15, 20, 22, 25, 28, 30, 35, 50. Найдите размах.

Задача 2. Новые случаи кори за неделю: 4, 6, 8, 10, 12, 14, 36. Найдите среднее и медиану. Кому верить?

Задача 3. Температура при внебольничной пневмонии: $n = 25$, среднее = 38.5°C , $s = 2.0^{\circ}\text{C}$. Найдите SEM и точный 95% ДИ ($df = 24$, $t = 2.064$).

Задача 4. Длительность лихорадки при ОРВИ (дни): 2, 3, 3, 4, 4, 5, 6, 7, 9. Найдите $Q1$, медиану, $Q3$ и IQR.

Ответы и решения — на следующих слайдах.



Задание 1 и Ответ для самопроверки

Инкубационный период гепатита А (дни): 15, 20, 22, 25, 28, 30, 35, 50. Найдите размах.

Ответ и решение:

Размах = $50 - 15 = 35$ дней. Значение 50 — вероятный выброс, в описании лучше добавить медиану.

Задание 2 и Ответ для самопроверки

Задача 2. Новые случаи кори за неделю: 4, 6, 8, 10, 12, 14, 36. Найдите среднее и медиану. Кому верить?

Ответ и решение:

Среднее = $90 / 7 \approx 12.9$; медиана = 10. Верить медиане: случай «36» тянет среднее вверх.

Задание 3 и Ответ для самопроверки

Задача 3. Температура при внебольничной пневмонии:
 $n = 25$, среднее $= 38.5$ °C, $s = 2.0$ °C. Найдите SEM и
точный 95% ДИ ($df = 24$, $t = 2.064$).

Ответ и решение:

$SEM = 2.0 / \sqrt{25} = 0.4$; ДИ $= 38.5 \pm 2.064 \times 0.4 = 38.5 \pm 0.83 \rightarrow$ от 37.7 до 39.3 °C.

Задание 4 и Ответ для самопроверки

Длительность лихорадки при ОРВИ (дни):

2, 3, 3, 4, 4, 5, 6, 7, 9.

Найдите Q1, медиану, Q3 и IQR.

Ответ и решение:

Q1 = 3-е значение = 3;

Медиана = 4;

Q3 = 7-е значение = 6;

IQR = 6 – 3 = 3 дня.

Шпаргалка изученных понятий

Показатель	Как считать	Зачем нужен
Размах	макс – мин	быстро, но боится выбросов
Среднее	$\Sigma x / n$	чувствительно к выбросам
SEM	$s / \sqrt{n}$	точность оценки среднего
95% ДИ (грубо)	среднее $\pm 2 \times \text{SEM}$	правило «большого пальца»
95% ДИ (точно)	среднее $\pm t \times \text{SEM}$	t — по таблице, $df = n - 1$
Медиана	середина упорядоченного ряда	устойчива к выбросам
Квартили	$Q1, Q3 \rightarrow \text{IQR} = Q3 - Q1$	ядро распределения

БЛАГОДАРЮ ЗА ВНИМАНИЕ

ИНФЕКЦИОННЫЕ БОЛЕЗНИ

