

Листки по ТВиМС
РЦ НИУ ВШЭ

Заикин Евгений Юрьевич

29 июля 2026 г.

Оглавление

1 Основные обозначения	5
I 2026 год	7
2 Основы статистики	9
2.1 Теоретические материалы	9
2.2 Примеры решения задач	11
2.3 Задачи для тренировки	12
2.4 Домашнее задание	16
3 Рассеивание данных - I	21
3.1 Теоретические материалы	21
3.2 Примеры решения задач	24
3.3 Задачи для тренировки	24
3.4 Домашнее задание	26
4 Рассеивание данных - II	29
4.1 Теоретические материалы	29
4.2 Примеры решения задач	31
4.3 Задачи для тренировки	33
4.4 Домашнее задание	34
5 Основы статистики. Точечные оценки параметров	37
5.1 Теоретический материал	37
5.2 Примеры решения задач	40
5.3 Задачи для тренировки	43
5.4 Домашнее задание	44
6 Линейная регрессия	47
6.1 Теоретические материалы	47
6.2 Примеры решения задач	51
6.3 Задачи для тренировки	53
6.4 Домашнее задание	53
7 Теорема Байеса в задачах	55
7.1 Теоретический материал	55
7.2 Задачи для тренировки	56
8 Проверка гипотез	61
8.1 Теоретические материалы	61
8.2 Задачи для тренировки	63

Ответы к задачам

67

Тема 1

Основные обозначения

Кванторы и логические символы

- $\forall$ - квантор всеобщности (*для любого*);
- $\exists$ - квантор существования (*существует*);
- $!$ - единственность;
- $\vee$ - дизъюнкция (*логическое или*);
- $\wedge$ - конъюнкция (*логическое и*).

Основные операции

- $A \cup B$ – объединение множеств A и B ;
- $A \cap B$ – пересечение множеств A и B ;
- $A \setminus B$ – разность множеств A и B ;

Основные множества

- $\emptyset$ – пустое множество;
- $\mathbb{N} = \{1, 2, 3, \dots\}$ – натуральные числа;
- $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ – целые числа;
- $\mathbb{Q} = \{\frac{m}{n} | m \in \mathbb{Z}, n \in \mathbb{N}\}$ – рациональные числа;
- $\mathbb{I}$ – иррациональные числа, непредставимые в виде рациональных;
- $\mathbb{R} = \mathbb{Q} \cup \mathbb{I}$ – вещественные (действительные) числа;
- $\mathbb{C} = \mathbb{R} \times i\mathbb{R}, i^2 = -1$. – комплексные числа;
- $\Omega = \{\omega_1, \dots, \omega_n\}$ – множество элементарных исходов некоторого эксперимента;
- 2^A – множество всех подмножеств (булеан) множества A .

Иные символы

- $\aleph$ – мощность счётного множества [алеф];

[Альфа]	α	A	[Йота]	ι	I	[Ро]	ρ	P
[Бета]	β	B	[Каппа]	κ	K	[Сигма]	σ	Σ
[Гамма]	γ	Γ	[Лямбда]	λ	Λ	[Тау]	τ	T
[Дельта]	δ	Δ	[Мю]	μ	M	[Ипсилон]	ν	Υ
[Эпсилон]	ε	E	[Ню]	ν	N	[Хи]	χ	X
[Дзета]	ζ	Z	[Кси]	ξ	Ξ	[Фи]	φ	Φ
[Эта]	η	H	[Омикрон]	o	O	[Пси]	ψ	Ψ
[Тета]	θ	Θ	[Пи]	π	Π	[Омега]	ω	Ω

Таблица 1.1: Греческий алфавит

Часть I
2026 год

Тема 2

ОСНОВЫ СТАТИСТИКИ

2.1 Теоретические материалы

Основные задачи статистики

Математическую статистику глобально интересуют два вопроса:

- Проверка статистических гипотез;
- Оценка некоторых параметров генеральной совокупности по параметрам выборки.

Например для первого пункта нам может быть интересно, есть ли зависимость между номером группы по математике и результатами независимой диагностики знаний обучающихся всех пяти групп, в предположении, что верна гипотеза

$$H_0 : \{ \text{чем меньше номер группы, тем выше балл за работу} \}.$$

Примером задачи второго типа служит понимание того, каким будет рейтинг кандидата на выборную должность после публикации некоторого пресс-релиза на основе полученного рейтинга от репрезентативной выборки фокус-группы.

Числовой набор и понятие выборки

Определение 2.1: Числовой набор

Конечную последовательность не обязательно различных чисел будем называть числовым набором. Обозначение: $a = (a_1, a_2, \dots, a_n)$ – набор из n чисел.

Как правило, числовой набор существует сам по себе, например это может быть набор оценок за триместр по предмету «теория вероятностей и математическая статистика» или рост учеников первой группы. С числовым набором можно выполнять линейные преобразования, например умножать каждый элемент набора a на некоторое число $\lambda \in \mathbb{R}$ или прибавлять к каждому элементу одно и то же число $k \in \mathbb{R}$. В таком случае удобно записывать получающиеся выборки коротко: λa и $a + k$. Для определённости:

$$\lambda a = (\lambda a_1, \lambda a_2, \dots, \lambda a_n) \text{ и } a + k = (a_1 + k, a_2 + k, \dots, a_n + k).$$

Говоря о выборках, мы говорим о представлении некоторой генеральной совокупности данных, то есть о таком числовом наборе, элементы которого случайно выбраны из большего числового набора, количество элементов которого столь велико, что становится затруднительным прямой анализ числовых характеристик. На первых парах на слова «выборка» и «числовой

набор» мы будем смотреть как на синонимы, но важно понимать, что выборка – гораздо более сложный предмет. Так выборки чрезвычайно полезны для анализа данных, которых очень много, например показатели роста всех людей на Земле генерируют колоссальный по количеству измерений числовой набор, как минимум насчитывающий 8028504258 элементов. Правда такой числовой набор для статистики интереса не представляет, хотя бы в силу того, что в него попадают и младенцы, и взрослые люди, чей рост сильно отличается.

Подкорректируем наш пример, оставив в наборе лишь те измерения, которые отвечают людям с возрастом от 24 до 50 лет. Теперь можно говорить о среднем или медианном росте, но измерений всё еще много. Поэтому, как правило, из генеральной совокупности случайным образом выбирается некоторый набор измерений, который и образует выборку. Если выполнить эту операцию методологически правильно, выборка действительно будет хорошо представлять исходный числовой набор.

Среднее арифметическое

*Средняя температура по больнице 36.6
с учетом гнойного отделения и морга*

Фольклор

Одной из самых часто используемых мер центральной тенденции является среднее арифметическое.

Определение 2.2: Среднее арифметическое

Для числового набора $a = (a_1, a_2, \dots, a_n)$ средним арифметическим называется число

$$\bar{a} = \frac{a_1 + a_2 + \dots + a_n}{n}.$$

Покажем, как ведёт себя среднее при линейном изменении каждого элемента числового набора.

Теорема 2.1: Линейность среднего арифметического

Пусть дан числовой набор a со средним $\bar{a}$, пусть так же $\lambda, k \in \mathbb{R}$. Тогда $\overline{\lambda a + k} = \lambda \bar{a} + k$.

Доказательство. Пусть $a = (a_1, \dots, a_n)$, тогда $\lambda a + k = (\lambda a_1 + k, \dots, \lambda a_n + k)$ и

$$\overline{\lambda a + k} = \frac{(\lambda a_1 + k) + \dots + (\lambda a_n + k)}{n} = \frac{\lambda(a_1 + \dots + a_n) + n \cdot k}{n} = \lambda \frac{a_1 + \dots + a_n}{n} + k = \lambda \bar{a} + k.$$

■

Медиана

Менее подверженной выбросам мерой центральной тенденции является медиана.

Определение 2.3: Медиана

Для упорядоченного по возрастанию числового набора $a = (a_1, a_2, \dots, a_n)$ $a_1 \leq a_2 \leq \dots \leq a_n$ медианой называется такое число, что половина из элементов набора не меньше него, а другая половина не больше.

Дадим ещё одно удобное определение медианы:

Определение 2.4: Медиана

Медиана набора чисел – это число, сумма расстояний от которого до всех чисел из набора минимальна.

Докажите эквивалентность приведённых определений медианы в качестве упражнения.

Мода

Иногда в статистических задачах приходится анализировать не числовые, а категориальные наборы, например набор любимых цветов учителей математики, ведущих в 7 классе: $s = (\text{красный, белый, оранжевый, зелёный, красный})$ или набор товаров, проданных за день в буфете $t = (\text{булочка с маком, сосиска в тесте, минипицца, сосиска в тесте})$.

Определение 2.5: Мода

Мода – одно или несколько значений набора, которое встречается наиболее часто.

Размах

Определение 2.6: Размах

Разница между наибольшим и наименьшим элементами числового набора называется размахом данного числового набора.

2.2 Примеры решения задач

Задача 1. В рамках проверки системы бюджетного контроля корпорации были проведены 10 измерений отклонений фактических расходов подразделений от плановых значений (в условных единицах). Результаты измерений представлены в таблице:

Номер измерения	1	2	3	4	5	6	7	8	9	10
Отклонение, у.е.	-2.2	0.4	8.7	-9.4	6.8	9.1	1.9	-3.5	5.3	-6.1

Для каждого измерения вычисляется параметр надёжности – модуль его отклонения от среднего арифметического. Измерение считается ненадёжным (подозрительным), если его параметр надёжности отличается от среднего значения параметра надёжности более чем в 1.6 раза.

Исключите из массива данных измерений ненадёжные. Найдите размах оставшихся измерений.

Решение.

Среднее арифметическое измерений:

$$\bar{x} = \frac{1}{10}(-2.2 + 0.4 + 8.7 - 9.4 + 6.8 + 9.1 + 1.9 - 3.5 + 5.3 - 6.1) = 1.1.$$

Параметры надёжности $d_i = |x_i - \bar{x}|$:

Номер	1	2	3	4	5	6	7	8	9	10
d_i	3.3	0.7	7.6	10.5	5.7	8.0	0.8	4.6	4.2	7.2

Средний параметр надёжности:

$$\bar{d} = \frac{1}{10}(3.3 + 0.7 + 7.6 + 10.5 + 5.7 + 8.0 + 0.8 + 4.6 + 4.2 + 7.2) = 5.26.$$

Границы интервала в 1.6 раза от $\bar{d}$:

$$\frac{\bar{d}}{1.6} = \frac{5.26}{1.6} = 3.2875, \quad 1.6\bar{d} = 1.6 \cdot 5.26 = 8.416.$$

Подозрительными считаются измерения, у которых $d_i \notin [3.2875; 8.416]$. Это измерения с номерами 2 ($d_2 = 0.7$), 4 ($d_4 = 10.5$) и 7 ($d_7 = 0.8$). Исключаем их. Оставшиеся значения: $-2.2, 8.7, 6.8, 9.1, -3.5, 5.3, -6.1$.

Размах — разность между наибольшим и наименьшим значениями:

$$9.1 - (-6.1) = 15.2.$$

Ответ: 15.2. ■

2.3 Задачи для тренировки

Задача 2.1. Пусть среднее $\bar{x} = 5$ для набора чисел x . Найдите $\overline{3x}$.

Задача 2.2. Покажите, что, если при поэлементном увеличении в k раз выборки (k не ноль и не единица) среднее не изменилось, то среднее в исходной выборке равнялось нулю.

Задача 2.3. Каждый элемент числового набора уменьшили на среднее этого набора. Найдите среднее арифметическое получившегося набора.

Задача 2.4. Среднее набора из девяти чисел было равно 8. К набору добавили число 5, чему стало равно среднее?

Задача 2.5. Среднее набора из восьми чисел было равно 9. К набору добавили число 4, на сколько изменилось среднее?

Задача 2.6. Средний рост класса, насчитывавшего 15 человек, составлял 161 см, к ним перевелся ученик и средний рост класса стал равен 160 см. Найдите рост новенького в классе.

Задача 2.7. В филиале «Сочно и восклицательный знак» работает 15 рядовых сотрудников, средняя зарплата которых составляет 33 т.р., и топ-менеджер филиала с зарплатой в 233 т.р. Какую среднюю зарплату по филиалу повесит на вакансию в филиал топ-менеджер и честно ли это?

Задача 2.8. Каждый элемент из набора чисел с нечётным числом элементов увеличили на 2. Как изменится медиана?

Задача 2.9. Где при решении предыдущей задачи нужно было воспользоваться нечетностью количества элементов?

Задача 2.10. Найдите медиану набора длин: 2 м 30 см, 250 мм, 0,02 км, 0,002 км, 2700 см, 2800 мм, 240 см.

Задача 2.11. В числовом наборе 100 чисел. Если выкинуть одно число, то медиана оставшихся чисел будет равна 78. Если выкинуть другое число, то медиана оставшихся чисел будет 66. Найдите медиану всего набора.

Задача 2.12. На нескольких розах сидели пчелы. В некоторый момент одна из пчел перелетела с цветка с максимальным количеством пчел на цветок с минимальным количеством пчел.

Найдите, как изменилось среднее количество пчел на одном цветке.

Задача 2.13. По результатам контрольной работы по математике выяснилось, что все ученики довольно хорошо подготовились, и более половины ребят получили оценку 5 из 5 возможных. Выберите верные утверждения:

- a) Среднее оценок не превосходит медианы;
- b) Нельзя точно сказать, чему равна медиана оценок;
- c) Нельзя точно сказать, чему равно среднее оценок;
- d) Нельзя точно сказать, чему равна мода оценок.

Задача 2.14. В парк аттракционов пришли пять классов школьников. Дети очень хотели на американские горки, но туда пускают детей с ростом не меньше 140 см. Известно, что:

- в 5-м «А» классе средний рост школьников равен 141 см;
- в 5-м «Б» медиана роста школьников равна 140 см;
- в 5-м «В» мода роста школьников равна 142 см;
- в 5-м «Г» у самого высокого школьника рост 145 см;
- в 5-м «Д» у самого низкого школьника рост 140 см.

Рост школьника может измеряться любым целым числом сантиметров. В каких классах хотя бы половина учеников сможет гарантированно попасть на аттракцион? Выберите все верные варианты ответа:

- a) 5 «А»;
- b) 5 «Б»;
- c) 5 «В»;
- d) 5 «Г»;
- e) 5 «Д».

Задача 2.15. При изучении иностранного языка класс делится на две группы. Ниже даны списки групп и полугодовые оценки учащегося. Может ли учительница английского языка перевести одного ученика из первой группы во вторую так, чтобы средний балл учащихся в обеих группах вырос?

№	1 группа	Отметка	№	2 группа	Отметка
1	Андреева	5	1	Алексеева	3
2	Борисов	3	2	Богданов	4
3	Васильева	5	3	Владимиров	5
4	Георгиев	3	4	Григорьева	2
5	Дмитриев	5	5	Давыдова	3
6	Евстигнеева	4	6	Евстахийев	2
7	Игнатов	3	7	Ильина	5
8	Кондратьев	4	8	Климова	4
9	Леонтьева	3	9	Лаврентьев	5
10	Миронов	4	10	Михайлова	3
11	Николаева	5			
12	Смирнов	5			

Задача 2.16. При испытании нового учебника по математике в эксперименте приняли участие примерно одинаковые по численности параллели седьмых классов из двух школ. В каждой

Учебник:	Экспериментальный	Обычный
Выполнили тест хорошо	30%	15%
Выполнили тест плохо	35%	20%
Доля тех, кто выполнил хорошо	0.46	0.43

Таблица 2.1: Результаты в первой школе

Учебник:	Экспериментальный	Обычный
Выполнили тест хорошо	20%	45%
Выполнили тест плохо	10%	25%
Доля тех, кто выполнил хорошо	0.67	0.64

Таблица 2.2: Результаты во второй школе

школе часть семиклассников училась по экспериментальному учебнику, а часть – по обычному старому учебнику. В конце учебного года был проведен итоговый тест, который показал, насколько хорошо семиклассники освоили учебный материал. В обеих школах доля тех, кто выполнил тест хорошо, оказалась выше в группе тех, кто учился по экспериментальному учебнику. Поэтому был сделан вывод, что новый учебник обеспечивает более высокое качество образования. Только Рассеянный Учёный выступил против нового учебника и ко всеобщему удивлению заявил, что те, кто учился по новому учебнику, сдали тест в среднем хуже, чем те, кто учился по старому. Кто прав?

Задача 2.17. В классе меньше, чем 35 учеников. Средний вес всех учеников равен 53.5 кг. Средний вес девочек 47 кг, а средний вес мальчиков – 60 кг. Может ли оказаться так, что самый легкий ученик в этом классе – мальчик, а самый тяжелый – девочка? Приведите пример, либо объясните, почему это невозможно.

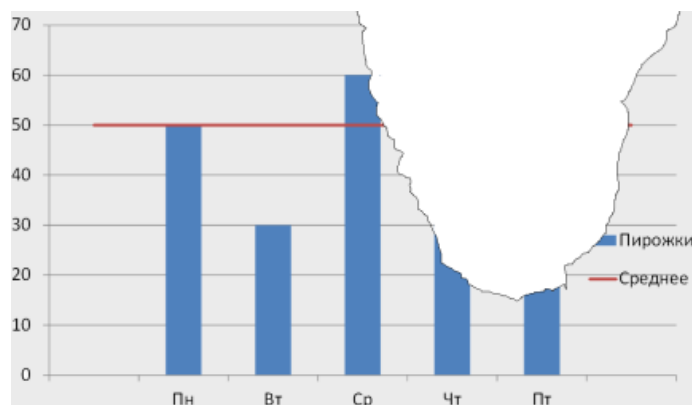
Задача 2.18. У Вали в шкатулке лежали пять разноцветных ленточек длиной 15, 20, 24, 26 и 30 см. Коля отрезал от некоторых (возможно, от всех) ленточек по кусочку. Средняя длина ленточек уменьшилась на 5 см, а медиана и размах длин не изменились. Чему теперь равны длины ленточек?

Задача 2.19. Гипермаркет, реализующий новогодние товары, состоит из трех отделов. В первом отделе представлены новогодние товары, цена каждого из которых меньше 100 руб. Средняя цена товаров в этом отделе равна 90 руб. Во втором отделе представлены новогодние товары, цена каждого из которых больше 100 руб. Средняя цена товаров в этом отделе равна 120 руб. Цена каждого товара в третьем отделе равна 100 руб. Средняя цена всех товаров в гипермаркете равна 110 руб., а общее число товаров равно 200. Все цены выражаются целым числом рублей.

- Может ли в первом отделе быть столько же товаров, сколько и во втором?
- Может ли в третьем отделе быть на 14 товаров больше чем во втором?
- Чему может равняться наибольшая возможная при этих условиях цена товара в этом гипермаркете?

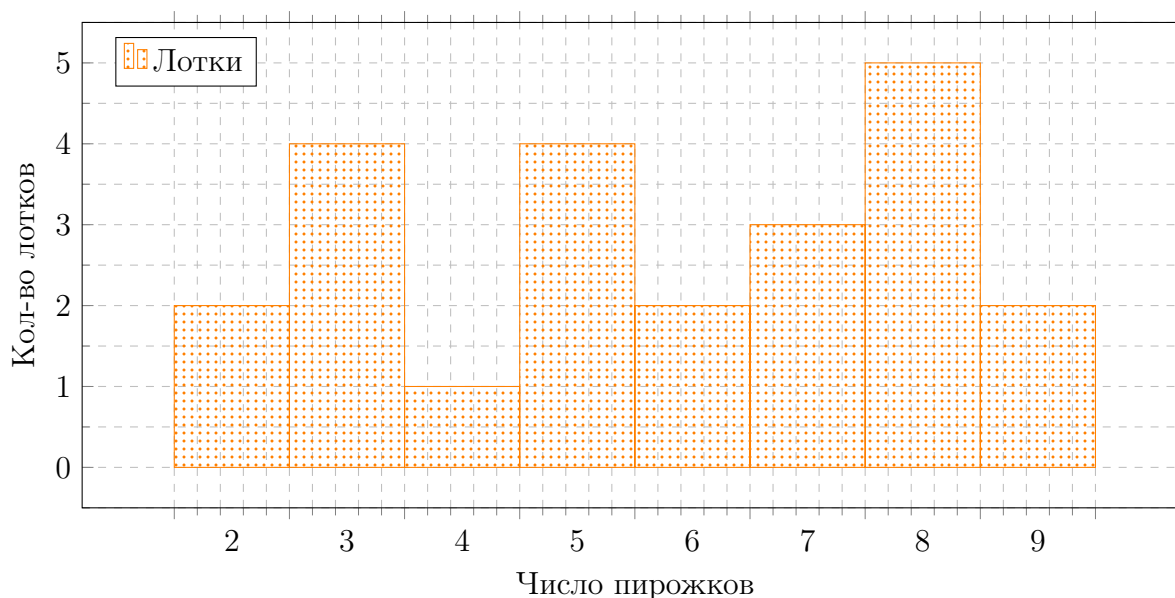
Задача 2.20. На диаграмме показано, сколько пирожков Вовочка съедал во все рабочие дни недели. Для наглядности среднее число съеденных пирожков отмечено горизонтальной линией.

Но случилось так, что кусок диаграммы Вовочка тоже откусил и съел. Известно лишь, что в четверг было съедено ровно 20% пирожков, съеденных за все пять дней. Сколько пирожков Вовочка съел в пятницу?



Задача 2.21. В сервисе такси «Ракета» рейтинг пассажира вычисляется как среднее арифметическое числа звезд, поставленных водителями пассажиру за несколько предыдущих поездок, более ранние поездки не учитываются. Водитель может поставить пассажиру от 1 до 5 звезд. Рейтинг не округляется. Ксения – курьер, и она регулярно пользуется такси «Ракета». До некоторых пор ее рейтинг был равен пяти. Однажды после поездки ее рейтинг снизился сразу до 4.94. Ксения очень расстроилась. Это было особенно неприятно, поскольку случилось в ее день рождения 15 ноября. Сможет ли Ксения, будучи предельно аккуратной, вежливой и пунктуальной пассажиркой, поднять свой рейтинг до наступления Нового года, если будет ездить на такси один раз каждый день?

Задача 2.22. В кафе 23 пластиковых лотка с пирожными. На диаграмме показано, как распределены пирожные. На горизонтальной прямой показано количество пирожных, а на вертикальной – сколько лотков содержит именно такое число пирожных. Найдите медиану величины «число пирожных в лотке».



Задача 2.23. Вася выписал в строку натуральные числа $a_1 < a_2 < \dots < a_{10}$. Оказалось, что среднее арифметическое этих чисел равно 25. Какое наименьшее значение может принимать число a_{10} ?

Задача 2.24. В классе присутствуют учитель и несколько учеников. Найти число учеников, если возраст учителя на 24 года больше среднего возраста учеников и на 20 лет больше среднего возраста всех присутствующих в классе.

2.4 Домашнее задание

Задача 2.25. В отделе работает 10 сотрудников. В ноябре самому высокооплачиваемому сотруднику повысили зарплату на 11400 рублей, а самому низкооплачиваемому понизили на 1600 рублей.

- а) Как изменилась медианная зарплата?
- б) Как изменился размах зарплат?
- с) Как изменилась средняя зарплата (среднее арифметическое)?

Задача 2.26. У студентов одного курса факультета инновационных исследований в среднем 22.425 однокурсниц. Сколько на этом курсе девушек?

Задача 2.27. Служить на подводной лодке может матрос, рост которого не превышает 168 см. Есть четыре команды А, Б, В и Г. Все матросы в этих командах хотят служить на подводной лодке и прошли строгий отбор. Остался последний отбор – по росту.

- В команде А средний рост матросов равен 166 см.
- В команде Б медиана роста матросов равна 167 см.
- В команде В самый высокий матрос имеет рост 169 см.
- В команде Г мода роста матросов равна 167 см.

В какой команде по крайней мере половина матросов точно может служить на подводной лодке?

Задача 2.28. Миша написал на доске число 1, а затем еще несколько натуральных чисел. Как только Миша пишет очередное число, Митя вычисляет медиану уже имеющегося набора чисел и записывает его себе в тетрадку. В некоторый момент в Митиной тетради записаны числа:

1; 2; 3; 2.5; 3; 2.5; 2; 2; 2; 2.5.

- а) Какое число Миша написал четвёртым?
- б) А какое – восьмым?

Задача 2.29. В классе находятся учитель и несколько учеников. Известно, что возраст учителя на 24 года больше среднего возраста учеников и на 20 лет больше среднего возраста всех присутствующих в классе. Сколько учеников находится в классе?

Задача 2.30. Среднее арифметическое десяти различных натуральных чисел равно 15. Найдите наибольшее значение наибольшего из этих чисел.

Задача 2.31. Выборка состоит из 3051 различного числа, причем все числа положительные и представимы в виде конечной десятичной дроби. Все числа из выборки упорядочили по возрастанию и разделили на 9 равных по величине групп, которые пронумеровали от 1 до 9. То есть первые 339 чисел упорядоченной выборки оказались в группе 1, следующие 339 чисел – в группе 2 и так далее. Выберите все верные утверждения:

- а) Среднее чисел в группе 3 больше среднего чисел в группе 2.
- б) Если одно число больше другого в 9 раз, то они обязательно окажутся в разных группах.
- с) Медиана выборки обязательно окажется в группе 5.
- д) Сумма двух чисел из группы 1 не может оказаться больше числа из группы 9.

е) Среди перечисленного выше нет верных ответов.

Задача 2.32. Приведите пример числового набора, который удовлетворяет трём следующим условиям.

- а) В наборе пять чисел;
- б) Среднее арифметическое набора больше, чем медиана;
- в) Из этого набора можно удалить число так, что среднее арифметическое оставшихся чисел станет меньше любой медианы оставшихся чисел.

Задача 2.33. Ученик пишет все самостоятельные и контрольные на 2. Других оценок он не получает. За неделю до выставления триместровой отметки он замечает среднее и успевает написать все работы еще раз, получая за каждую новую 5. Как за эту неделю изменились: размах, медиана, мода и среднее набора?

Задача 2.34. В первой четверти у Васи было пять оценок по математике, больше всего среди них пятёрок. При этом оказалось, что медиана всех оценок равна 4, а среднее арифметическое 3.8. Какие оценки могли быть у Васи?

Задача 2.35. В таблице ниже представлено распределение ответов школьников на вопрос «Как изменилась ваша успеваемость в этом учебном году по сравнению с предыдущим?». Все ученики отвечали на вопрос правдиво, таким образом, представленные числа отражают реальное изменение успеваемости.

Ответ	Частота
Значительно улучшилась	16
Немного улучшилась	8
Не изменилась	12
Немного ухудшилась	4
Сильно ухудшилась	3

Выберите все верные варианты ответа:

- а) Успеваемость улучшилась более чем у половины школьников;
- б) Успеваемость не улучшилась более чем у половины школьников;
- в) У медианного школьника успеваемость не изменилась
- г) У медианного школьника успеваемость немного улучшилась;
- е) Среди перечисленного выше нет верных ответов.

Задача 2.36. Из цифр от 1 до 9 составлены три однозначных и три двузначных числа, причем цифры не повторяются. Найдите наименьшее возможное среднее арифметическое получившегося набора чисел.

Задача 2.37. В понедельник 15 ноября Митя разместил на своей странице в соцсети фотографию своего пса Харви. Диаграмма показывает, сколько лайков это фото получило за 10 дней, но подписи на вертикальной оси отсутствуют. Известно, что в четверг 18 ноября лайков было на четыре меньше, чем накануне. Среднее количество лайков за все дни оказалось 19. Сколько лайков получило фото в воскресенье 21 ноября?

Задача 2.38. Из 40 последовательных нечетных чисел $1, 3, 5, \dots, 79$ выбрали 7 различных чисел в порядке возрастания и вычислили их среднее арифметическое $\bar{x}$ и медиану m .

- а) Может ли разность $\bar{x} - m$ равняться $\frac{2}{7}$?
- б) Может ли разность $\bar{x} - m$ равняться $\frac{3}{7}$?

Задача 2.39. На доске было написано 20 натуральных чисел (не обязательно различных), каждое из которых не превосходит 40. Вместо некоторых из чисел (возможно, одного) на доске написали числа, меньшие первоначальных на единицу. Числа, которые после этого оказались равными 0, с доски стёрли.

- а) Могло ли оказаться так, что среднее арифметическое чисел на доске увеличилось?
- б) Среднее арифметическое первоначально написанных чисел равнялось 27. Могло ли среднее арифметическое оставшихся на доске чисел оказаться равным 34?
- в) Среднее арифметическое первоначально написанных чисел равнялось 27. Найдите наибольшее возможное значение среднего арифметического чисел, которые остались на доске.

Задача 2.40. На проекте «Мисс года – 2026» выступление каждой участницы оценивают шесть судей. При этом каждый судья выставляет оценку – целое число баллов от 0 до 10 включительно. Известно, что за выступление Изольды Кабановой все члены жюри выставили различные оценки. По старой системе оценивания итоговый балл за выступление определяется как среднее арифметическое всех оценок судей. По новой системе оценивания итоговый балл вычисляется следующим образом: отбрасываются наименьшая и наибольшая оценки, и считается среднее арифметическое четырёх оставшихся оценок.

- а) Могут ли итоговые баллы, вычисленные по старой и новой системам оценивания, оказаться одинаковыми?
- б) Может ли разность итоговых баллов, вычисленных по старой и новой системам оценивания, оказаться равной $\frac{1}{8}$?
- в) Найдите наибольшее возможное значение разности итоговых баллов, вычисленных по старой и новой системам оценивания.

Задача 2.41. Годовую отметку учитель обычно выводит, усредняя отметки в два этапа. Сначала текущие отметки усредняют и при необходимости округляют до целого, чтобы поставить четвертную отметку. В конце года усредняют и округляют четвертные отметки, чтобы получить годовую. Если среднее значение получается полуцелым, то принято округлять в большую сторону. В таблице показан пример.

В этом примере итоговая годовая отметка 4 больше, чем среднее значение всех текущих отметок, но ненамного (на 0.24). Возникает масса вопросов, связанных с тем, как и насколько двойное усреднение с округлением искажает истинное среднее значение. Может ли случиться так, что отличие окажется больше, чем 0.5? То есть, например, может ли быть годовая четвертка при среднем значении всех текущих меньше, чем 3.5? Может ли случиться так, что годовая отметка окажется ниже, чем среднее всех текущих? Каково наибольшее возможное отличие годовой от средних текущих в ту и в другую сторону? Попробуйте привести примеры разных экзотических ситуаций или придумать рассуждения, доказывающие, что какие-то комбинации невозможны. Напишите серьезное и вдумчивое эссе на эту животрепещущую школьную тему.

	Оценки						Усреднение	Округление
I четверть	3	3	4	3	5	3	$\frac{3+3+4+3+5+3}{6} = 3.5$	4
II четверть	4	3	4	4			$\frac{4+3+4+4}{4} = 3.75$	4
III четверть	4	3	5	5	4	5	$\frac{4+3+5+5+4+5}{6} \approx 4.33$	4
IV четверть	4	3	4	2	4		$\frac{4+3+4+2+4}{5} = 3.4$	3
Годовая							$\frac{4+4+4+3}{4} = 3.75$	4
Среднее							$\frac{79}{21} \approx 3.76$	

Тема 3

Рассеивание данных - I

*Часто средства оправдывают цели:
Цели порождают метод, а метод
выживает, даже когда разрушаются
структуры, бывшие ранее целью.*

Алан Перлис

3.1 Теоретические материалы

Определение 3.1: Дисперсия

Дисперсией числового набора $a = (a_1, a_2, \dots, a_n)$ называется среднее суммы квадратов отклонений:

$$\mathbb{D}(a) = \frac{(\bar{a} - a_1)^2 + (\bar{a} - a_2)^2 + (\bar{a} - a_3)^2 + \dots + (\bar{a} - a_n)^2}{n}.$$

Теорема 3.1: Свойства дисперсии

Дисперсия числового ряда a имеет следующие свойства:

- a) $\mathbb{D}(a + k) = \mathbb{D}(a)$, $k \in \mathbb{R}$;
- b) $\mathbb{D}(ka) = k^2 \mathbb{D}(a)$, $k \in \mathbb{R}$.

Теорема 3.2: Свойство дисперсии

Дисперсия может быть вычислена как $\mathbb{D}(a) = \overline{a^2} - (\bar{a})^2$.

Сама дисперсия пригождается часто лишь как инструмент для более точного показателя – стандартного отклонения.

Стандартное отклонение

Определение 3.2: Стандартное отклонение

Стандартным отклонением набора чисел называется квадратный корень из дисперсии этого набора.

$$\sigma(a) = \sqrt{\mathbb{D}(a)}.$$

Теорема 3.3: Нулевое стандартное отклонение

Если стандартное отклонение некоторого набора равно нулю, то все его элементы равны.

Теорема 3.4: Свойства стандартного отклонения

Стандартное отклонение числового набора при сдвиге не меняется:

$$\sigma(a + \lambda) = \sigma(a), \quad \lambda \in \mathbb{R}.$$

Если каждый элемент набора умножить на $\lambda \in \mathbb{R}$, его стандартное отклонение умножится на $|\lambda|$:

$$\sigma(\lambda a) = |\lambda| \cdot \sigma(a).$$

Зачем нужно стандартное отклонение

Предположим, вы собрали некоторую статистику: набор измерений по какому-то параметру для большого числа данных, то есть генеральную совокупность. Как оценить, что для полученной совокупности данных можно считать нормой, что превышением, а что недостатком?

Прежде всего важно помнить, что статистика, не смотря на всю кажущуюся точность похожа на гуманитарную дисциплину. Например утренняя температура «жаворонков» составляет $36.7 \pm 0.11^\circ\text{C}$, а «сов» – $36.4 \pm 0.21^\circ\text{C}$. Как видно, показатель в 36.6°C , к которому многие привыкли ниже среднего для «жаворонков» и выше среднего для «сов». Кроме того, бывают исключения, в данном случае люди, для которых их личная норма отличается от нормы остальных, даже интервальной. К сожалению для последних, современная наука часто может оставить их «за бортом».

Но давайте временно забудем о таких сложных этических вопросах и вернемся к математике. Пусть наша генеральная совокупность описывает, например, население городов. Можно ли выбрать некоторый показатель, который позволил бы нам навесить на города ярлыки «крупный», «стандартный» (если хотите, «типичным») и «маленький»?

Рассмотрим данные о населении городов Подмосковья. Удалим два очень больших и одно очень маленькое значение (население городов Балашиха, Подольск и Верея). Найдём среднее и стандартное отклонение после удаления этих трёх городов: $\bar{x} = 67.12$ тыс. чел., $\sigma(x) = 55.97$ тыс. чел. Если отступить на одно стандартное отклонение от среднего влево и вправо, получится интервал от 11.15 до 123.09 тыс. чел. В таком случае будем считать, что на интервале $(\bar{x} - \sigma(x); \bar{x} + \sigma(x))$ находятся «стандартные» по населению города. Такое правило основано на одном стандартном отклонении от среднего. В качестве решающего критерия часто берут интервал в два или полтора стандартных отклонения, в старших классах вы докажете так называемое «правило трёх сигм»: не менее 89% измерений попадают в интервал $(\bar{x} - 3\sigma(x); \bar{x} + 3\sigma(x))$.

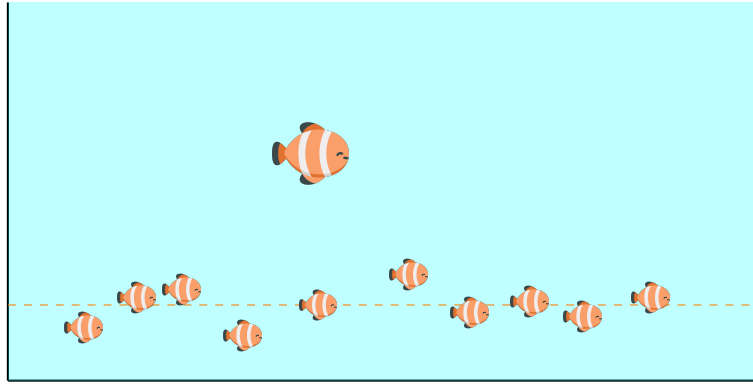
Такие правила позволяют найти в выборке так называемые «выбросы» – нехарактерные относительно остальных элементов выборки значения. Некоторые выбросы должны быть исключены из набора данных, так как их причинами являются ошибки и технические неполадки; другие выбросы необходимо оставить в наборе данных. Если, например, выброс не является результатом ошибки и/или дает новое понимание тестируемого явления, то его нужно оставить в

№ п/п	Город	Население, тыс. чел.	№ п/п	Город	Население, тыс. чел.
1	Апрелевка	29.1	38	Лыткарино	58.6
2	Балашиха	490.0	39	Люберцы	207.3
3	Бронницы	22.6	40	Можайск	30.0
4	Верея	5.0	41	Мытищи	222.7
5	Видное	71.4	42	Наро-Фоминск	63.0
6	Волоколамск	19.2	43	Ногинск	102.3
7	Воскресенск	93.2	44	Одинцово	137.5
8	Высоковский	10.4	45	Озёры	25.0
9	Голицыно	17.4	46	Орехово-Зуево	118.0
10	Дедовск	30.1	47	Павловский Посад	63.9
11	Дзержинский	56.3	48	Пересвет	13.6
12	Дмитров	68.7	49	Подольск	304.2
13	Долгопрудный	112.0	50	Протвино	35.8
14	Домодедово	133.5	51	Пушкино	106.8
15	Дрезна	11.3	52	Пушино	20.8
16	Дубна	75.0	53	Раменское	119.9
17	Егорьевск	72.1	54	Реутов	107.0
18	Жуковский	107.9	55	Рошаль	20.2
19	Зарайск	22.9	56	Руза	13.0
20	Звенигород	21.9	57	Сергиев Посад	102.0
21	Ивантеевка	79.3	58	Серпухов	124.9
22	Истра	33.5	59	Солнечногорск	51.6
23	Кашира	48.0	60	Старая Купавна	22.5
24	Клин	79.4	61	Ступино	66.0
25	Коломна	141.1	62	Талдом	12.7
26	Королёв	224.5	63	Фрязино	60.0
27	Котельники	46.8	64	Химки	254.7
28	Красноармейск	26.5	65	Хотьково	21.2
29	Красногорск	171.8	66	Черноголовка	21.1
30	Краснозаводск	12.9	67	Чехов	71.9
31	Краснознаменск	42.5	68	Шатура	32.4
32	Кубинка	20.1	69	Щёлково	124.8
33	Куrowsкое	20.6	70	Электрогорск	22.8
34	Ликино-Дулёво	29.0	71	Электросталь	157.4
35	Лобня	89.3	72	Электроугли	20.7
36	Лосино-Петровский	25.3	73	Яхрома	14.3
37	Луховицы	30.3			

Таблица 3.1: Население городов Подмосквья

наборе данных. Научные эксперименты особенно чувствительны к выбросам – исключив выброс по ошибке, вы можете пропустить некоторую новую тенденцию или открытие.

Например, мы разрабатываем новый препарат для увеличения размера рыб в рыбном хозяйстве. Мы будем использовать набор данных (71, 70, 73, 70, 70, 69, 70, 72, 71, 300, 71, 69) (см. рис. ??), каждое значение будет представлять массу рыбы (в граммах) после приема экспериментального препарата.



Размеры рыбок

Другими словами, первый препарат приводит к увеличению массы рыбы до 71 г, второй препарат – до 70 г и так далее. В этой ситуации 300 – это значительный выброс, но мы не должны исключать его; если предположить, что не было ошибок измерения, то такой выброс – это значительный успех в эксперименте. Препарат, который увеличил вес рыбы до 300 граммов, действует значительно лучше других препаратов; таким образом, 300 – это самое важное значение в наборе данных.

3.2 Примеры решения задач

3.3 Задачи для тренировки

Задача 1. Для выборки $a = (5, 13, 8, 12, 12)$ найдите:

- а) Среднее значение $\bar{a}$; б) Выборку a^2 ; в) $\mathbb{D}(a)$.

Задача 2. Про набор чисел x известно, что $D(x) = 0$. Приведите пример такого набора.

Задача 3. Каждый элемент числового набора умножили на $-\frac{1}{5}$. Какой станет дисперсия этого набора, если до умножения она была равна 50?

Задача 4. Пусть известно, что $\mathbb{D}(a) = 5$. Найдите:

- а) $\mathbb{D}(2a)$; б) $\mathbb{D}(3a - 1)$; в) $\mathbb{D}(4 - 17a)$; г) $\mathbb{D}(0a)$.

Задача 5. Известно, что сумма всех отклонений кроме одного от среднего значения по выборке равна 3. Найдите отклонение от среднего для оставшегося элемента выборки.

Задача 6. В наборе $-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$ замените одно число двумя другими целыми числами так, чтобы дисперсия набора и его среднее не изменились.

Задача 7. Вычислите дисперсию числового набора $\{11, 22, 33, 44, 55, 66, 77\}$.

Задача 8. Вычислите дисперсию числового набора $\{17, 32, 47, 62, 77\}$.

Задача 9. Вычислите стандартное отклонение набора чисел $a = (3, 4, 5, 3, 3)$.

Задача 10. Каждый элемент набора чисел со стандартным отклонением равным 5.8 увеличили в 1.5 раза. Чему равно стандартное отклонение полученного набора?

Задача 11. Пусть известно, что $\sigma(x) = 1.23$. Найдите $\sigma(-2x)$.

Hint В следующей задаче предполагается, что все знают теорему Пифагора: В прямоугольном треугольнике сумма квадратов катетов равна квадрату гипотенузы.

Задача 12. Оцените сверху дисперсию набора длин сторон прямоугольного треугольника с гипотенузой 6.

Задача 13. Длина гипотенузы прямоугольного треугольника равна 3. Рассеянный Учёный вычислил дисперсию длин сторон этого треугольника и нашёл, что она равняется 2. Не ошибся ли он в расчетах?

Задача 14. Числа Фибоначчи f_n определяются следующим образом:

$$\bullet f_0 = 1, \quad \bullet f_1 = 1, \quad \bullet f_n = f_{n-1} + f_{n-2};$$

для $n \in \mathbb{N}, n \geq 2$, например $f_2 = 1 + 1 = 2$, $f_3 = 1 + 2 = 3$ и т.д. Вычислите дисперсию первых k чисел Фибоначчи, где:

$$\text{a) } k = 2, \quad \text{b) } k = 3, \quad \text{c) } k = 4, \quad \text{d) } k = 5.$$

Задача 15. Докажите, что в предыдущей задаче с ростом k дисперсия растёт.

Задача 16. В числовом наборе чисел, среди которых есть числа 0 и 1. Какова наименьшая возможная дисперсия такого набора чисел? Каким для этого должен быть набор?

Задача 17. Дан набор чисел $a = (4, 3, 2, 1, 9, 7, 2, 7, 1, 4)$. Постройте по этим данным интервал, получающийся отступом от среднего на одно стандартное отклонение.

Задача 18. Для некоторой выборки с дисперсией 1024 четыре раза повторили следующую операцию: к каждому числу добавили 1, а затем каждое число разделили на 2. Чему равняется стандартное отклонение полученного в итоге набора?

Задача 19. Вычислите дисперсию числового набора

$$s = (57^{57}, 57^{57}, 57^{57}, 57^{57} + 1, 57^{57} + 2).$$

Задача 20. Вычислите дисперсию набора:

$$s = (77^7, 77^7 + 1, 77^7 - 1, 77^7, 77^7 + 2).$$

Задача 21. В наборе данных стёрли все числа кроме одного, которое равно π . Чему могли быть равны остальные 9 чисел, если известно, что стандартное отклонение набора равно нулю?

Задача 22. Из некоторого набора чисел x сделали набор $y = \frac{x - \bar{x}}{\sigma(x)}$. Найдите среднее и стандартное отклонение полученного набора.

Задача 23. Собраны результаты средних баллов по ТВиМС группы из 20 человек. Укажите порядковые номера измерений, попадающих в интервал отклонений от среднего на одно стандартное отклонение.

1	2	3	4	5	6	7	8	9	10
4.00	4.00	3.67	4.33	3.00	4.00	3.33	4.00	3.75	2.67
11	12	13	14	15	16	17	18	19	20
3.00	3.67	2.00	4.00	3.75	2.00	2.67	3.67	4.33	5.00

Задача 24. Ниже представлены результаты измерения роста девушек в возрасте 20 – 21 года. Найдите для данной выборки:

Задача 9. Каждый элемент набора чисел со стандартным отклонением равным 5.8 уменьшили на 3. Чему равно стандартное отклонение полученного набора?

Задача 10. В рамках примера в таблице ?? назовём город «очень крупным», если его население больше чем $\bar{x} + 2\sigma(x)$. Найдите по таблице очень крупные города.

Задача 11. По таблице выберите ингредиент салата Оливье, для которого абсолютное отклонение числа килокалорий от среднего для салата максимально.

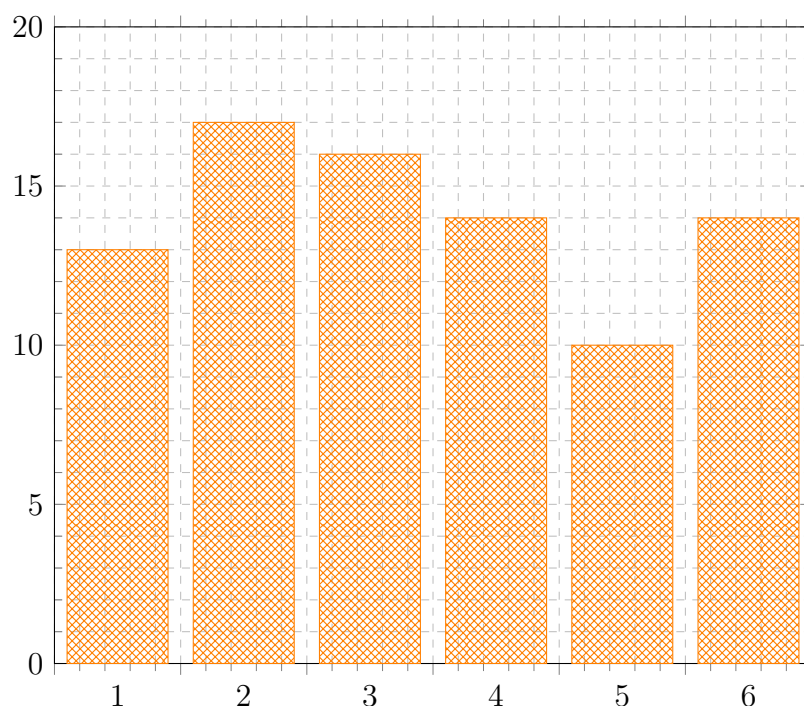
Продукт	Кол-во	Вес, г	На 100 граммов продукта			
			Белки	Жиры	Углеводы	Килокалории
Колбаса докторская	50 г	50	13.4	22.9	0	257
Картофель	0.4 шт	40	2	0.3	16.5	80
Морковь	0.2 шт	20	1.3	0.1	6.3	29
Огурец	0.2 шт	20	0.7	0	3.1	15
Зелёный горошек	30 г	30	5.5	0.7	7.4	74
Куриные яйца	0.4 шт	20	12.7	11.1	0.6	153
Майонез	20 г	20	0.3	67	2.6	614.6

Задача 12. Вычислите дисперсию числового набора $s = (58^{58}, 58^{58}, 58^{58}, 58^{58} + 1, 58^{58} + 2)$.

Задача 13. В таблице ниже представлены измерения высот высочайших гор планеты. Найдите стандартное отклонение данной выборки и горы, которые не попадают в интервал для одного отклонения на стандартное отклонение от среднего данной выборки.

Джомолунгма	Чогори	Улугмузтаг	Тиричмир	Пик Конгур
8848	8614	7723	7690	7649
Гонгга	Пик Победы	Ньэнченгангла	Аконкагуа	Койо Зом
7590	7439	7162	6961	6872

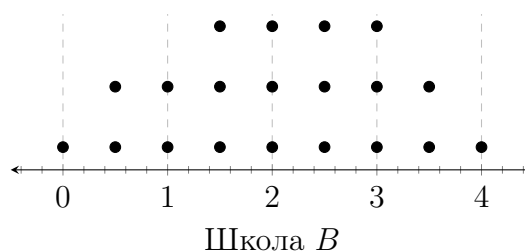
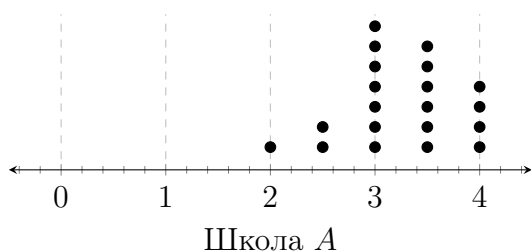
Задача 14. В каждой группе по математике решили играть в Тайного Санту. Число игроков в каждой из групп представлено в диаграмме. Найдите группу или группы, число игроков в которых отклоняется от среднего числа игроков по группам хотя бы на одно стандартное отклонение.



Задача 15. Каждый элемент набора чисел с дисперсией 1 заменили противоположным числом, а затем увеличили на 5. На сколько изменилось стандартное отклонение данного набора?

Задача 16. На точечных графиках показаны средние баллы выпускников двух школ в одном районе, школы A и школы B . Каждая точка соответствует 5 ученикам, и в каждой школе по 100 выпускников. Что из перечисленного ниже правильно сравнивает стандартное отклонение оценок в каждом из классов?

- a) Стандартное отклонение оценок в школе A меньше.
- b) Стандартное отклонение оценок в школе B меньше.
- c) Стандартное отклонение оценок в школе A и школе B одинаково.
- d) Из предоставленной информации невозможно определить взаимосвязь.



Задача 17. Даны 100 чисел. К каждому из них прибавили 2. Сумма квадратов чисел не изменилась. К каждому полученному числу еще раз прибавили число 2. Как изменилась сумма квадратов теперь?

Задача 18 (Африканская саванна). Предположим, что зоологами в ходе исследования фиксировалась численность (размер) прайдов в африканской саванне. Прайдом в зоологии называется семейная стая львов, состоящая из нескольких особей (численность прайда может достигать трёх-четырёх десятков животных). Зоологами указывалось количество животных в каждом прайде. Найдите среднее число особей в прайде и стандартное отклонение этого числа. Выборка состоит из 27 наблюдений (прайдов, охваченных исследованием):

Численность прайда	1	3	5	6	7	8	9	11	13	20	Всего
Прайдов с такой численностью	2	1	3	2	4	4	5	3	2	1	$n = 27$

Таблица 3.2: Численность прайдов

Тема 4

Рассеивание данных - II

4.1 Теоретические материалы

Определение 4.1: Квантиль

Квантилем уровня $\alpha \in (0,1)$ выборки размера n называется такое число x_α , что доля равная α элементов этой выборки не больше x_α и доля, равная $1 - \alpha$ – не больше.

У этого определения есть часто используемые частные случаи:

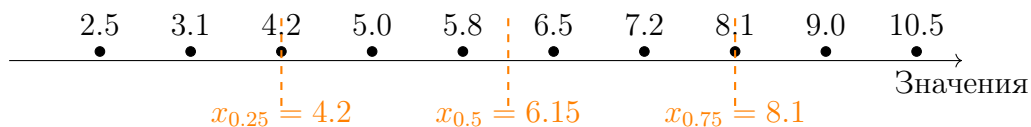
Определение 4.2: Медиана

Медианой называется квантиль уровня 0.5.

Обратите внимание на то, что по определению медиана, как и остальные квантили вовсе не обязаны определяться однозначно. Между тем мы будем считать, что для выборки X из $2k + 1$ элемента медиана есть в точности $k + 1$ элемент в вариационном ряду (упорядоченной выборке), для выборки же X из $2k$ элементов положим медиану равной среднему арифметическому центральных членов: $x_{0.5} = \frac{x_k + x_{k+1}}{2}$.

Определение 4.3: Квартили

Квартелями (Q_1 первым, Q_2 вторым и Q_3 третьим называются, соответственно, квантили $x_{\frac{1}{4}}$, $x_{\frac{1}{2}}$, $x_{\frac{3}{4}}$).



Очевидно, что второй квартиль есть медиана.

Определение 4.4: Дециль

Дециль – это квантиль уровня $\frac{1}{10}$.

Определение 4.5: Перцентиль

Перцентиль – это квантиль уровня $\frac{1}{100}$.

Определение 4.6: Межквартильный размах

Межквартильным размахом называется величина $IQR = Q_3 - Q_1$.

о указать границы характерных значений выборки. Такие границы бывают полезны для отдельного изучения тех наблюдений, которые не попали в эти границы.

Определение 4.7: Область характерных значений

Областью характерных значений выборки называется промежуток

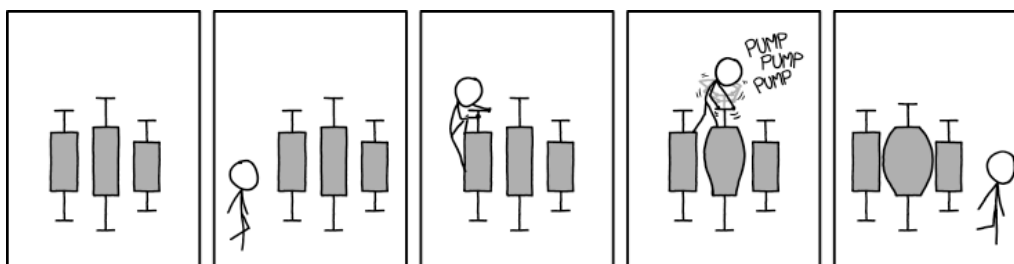
$$(x_{0.25} - 1.5IQR; x_{0.75} + 1.5IQR).$$

Часто для дополнительного уточнения вводят и такое определение:

Определение 4.8: Существенно отклоняющиеся варианты

Варианта называется существенно отклоняющейся, если не попадает в интервал:

$$(x_{0.25} - 3IQR; x_{0.75} + 3IQR).$$

Ящик с усами

Ящик с усами – график, позволяющий быстро проводить визуальный анализ данных.

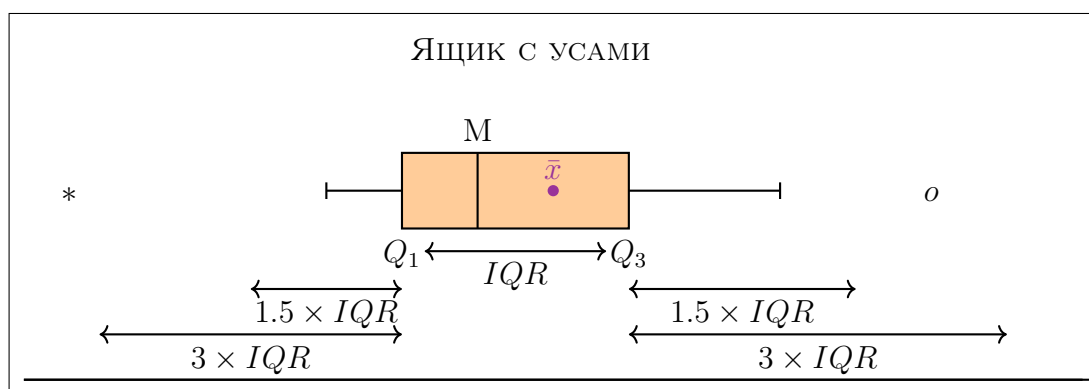


Диаграмма «Ящик с усами» строится в следующей последовательности операций (шагов):

1. Рисуются прямоугольник (так называемый «ящик», коробка). Его нижней планкой будет первый квартиль выборки Q_1 , а верхней – третий квартиль Q_3 . Высота прямоугольника соответствует квартильному размаху $IQR = Q_3 - Q_1$;
2. Внутри прямоугольника в соответствующем месте проводится полоска – медиана. Она может совпадать с центром «ящика», располагаться ближе к одному из квартилей или касаться одной из его границ;

3. Рисуются «отростки» (так называемые усы), направления которых зависят от того, есть ли в выборке нетипичные значения:
- Если в выборке нет нетипичных значений (т. е. все наблюдения в выборке охватываются, «накрываются» интервалом $[Q_1 - 1,5IRQ; Q_3 + 1,5IRQ]$), то «усы» отходят от Q_3 вверх и от Q_1 вниз до точек X_{min} (фактический минимум) и X_{max} (фактический максимум) соответственно;
 - Если в выборке присутствуют нетипичные значения («подозрительные» и/или «выбросы»), то «усы» отходят от Q_3 вверх и от Q_1 вниз до точек X_{max} (максимальное из тех значений выборки, которые принадлежат диапазону $[Q_3; Q_3 + 1.5IRQ]$) и X_{min} (минимальное из тех значений выборки, которые принадлежат диапазону $[Q_1 - 1.5IRQ; Q_1]$);
4. На диаграмме особым образом помечаются нетипичные наблюдения выборки: точками (или кружочками) – подозрительные наблюдения, звёздочками – выбросы.

4.2 Примеры решения задач

Задача 1. Найдите квартили выборки $x = (40, 25, 60, 15, 10, 10, 20, 40, 30, 35, 40, 85)$.

Решение.

Упорядочим нашу выборку:

$$X = (10, 10, 15, 20, 25, 30, 35, 40, 40, 40, 60, 85).$$

Всего в выборке $n = 12$ элементов, значит каждая квартиль отсекает 3 элемента. Вторая квартиль, то есть медиана, $Q_2 = x_{\frac{1}{2}}$ – любое число из промежутка $(30, 35)$. Для определённости возьмем среднее арифметическое границ: $Q_2 = x_{\frac{1}{2}} = \frac{30 + 35}{2} = 32.5$.

Аналогично находим квартили $Q_1 = x_{\frac{1}{4}}$ и $Q_3 = x_{\frac{3}{4}}$, по совместительству являющиеся медианами подвыборок $(10, 10, 15, 20, 25, 30)$ и $(35, 40, 40, 40, 60, 85)$.

Имеем $Q_1 = 17.5$, $Q_3 = 40$. ■

Задача 2 (Рестораны). Рассмотрим финансовые результаты деятельности двух ресторанов быстрого питания – А и Б. Пусть нас интересует такая величина, как сумма чека в одном заказе посетителя. Имеются две выборки данных:

- ресторан А, X (руб.): 50, 600, 500, 400, 300, 400, 500, 450, 500 – суммы чеков;
- ресторан Б, X (руб.): 700, 200, 400, 500, 200, 500, 700, 500 – суммы чеков;

Постройте диаграммы «Ящик с усами» для каждого ресторана.

Решение.

Проведём подготовительные расчёты, требующиеся для визуализации данных.

- Рассмотрим первую выборку, упорядочим её значения от минимума к максимуму:

$$X_a = 50, 300, 400, 400, 450, 500, 500, 500, 600;$$

- Вычислим медиану, квартили и квартильный размах: $M = Q_2 = 450$, $Q_1 = 400$, $Q_3 = 500$, $IRQ = 500 - 400 = 100$.
- Найдём область характерных (типичных) для этой выборки значений: $X_{min} = 400 - 1,5 \cdot 100 = 400 - 150 = 250$; $X_{max} = 500 + 1,5 \cdot 100 = 500 + 150 = 650$. Получим интервал $[250; 650]$.

4. Теперь посмотрим, есть ли в выборке наблюдения, не вошедшие в этот интервал, т. е. являющиеся нетипичными. Да, значение 50 не принадлежит области характерных значений.
5. Определим, к какой категории нетипичных значений – «подозрительным» или «выбросам» – относится значение «50 руб.». Для этого рассчитаем более широкий диапазон, сформированный точками, отстоящими от квартилей на $3 \times IRQ$. Получим:

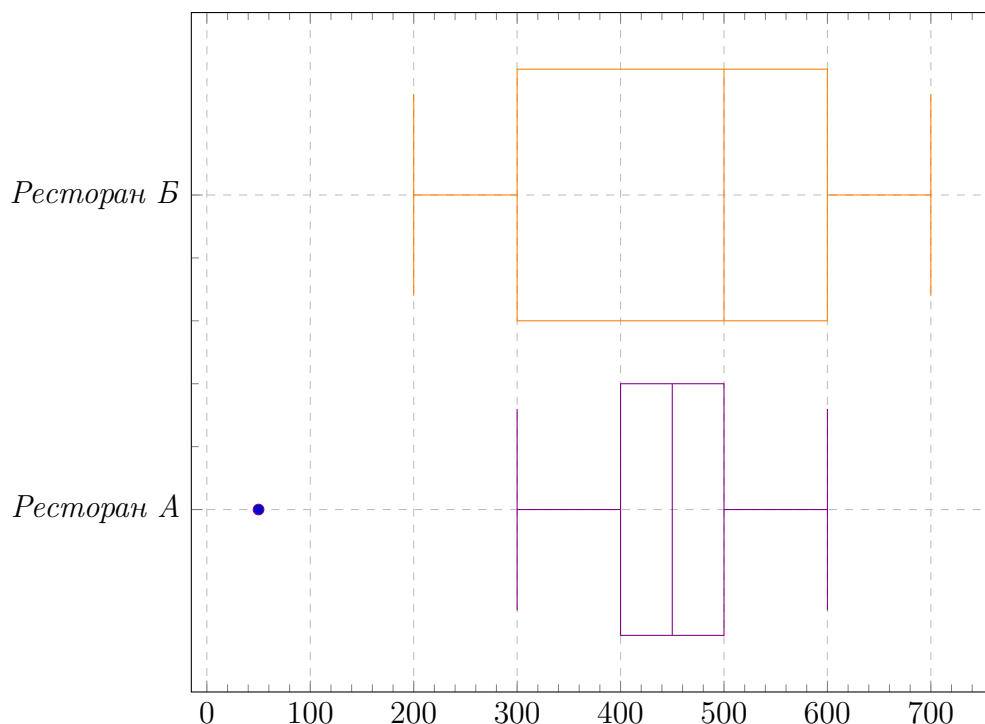
$$[400 - 3 \cdot 100; 500 + 3 \cdot 100] = [100; 800].$$

Полученный интервал не захватывает значение 50. Следовательно, значение 50 – «выброс».

Теперь все сделанные операции повторим для второй выборки (данные по ресторану Б).

1. Сформируем упорядоченный ряд: X_b : 200, 200, 400, 500, 500, 500, 700, 700;
2. $M = Q_2 = 500$, $Q_1 = 300$, $Q_3 = 600$, $IRQ = 600 - 300 = 300$.
3. Область характерных значений: $X_{min} = 300 - 1,5 \cdot 300 = 300 - 450 = -150$. Значение -150 можно округлить до нуля, так как сумма заказа в чеке не может быть отрицательной. $X_{max} = 600 + 1,5 \cdot 300 = 600 + 450 = 1050$. Получен интервал $[0; 1050]$.
4. В этой выборке отсутствуют нетипичные значения наблюдений.

Сравнение чеков в ресторанах



Построим на одном рисунке диаграммы «Ящик с усами» для ресторанов А и Б. Эти изображения дают довольно исчерпывающую информацию о структуре данных в рассматриваемых выборках (ресторан А и ресторан Б). Мы нарисовали «Ящики с усами» бок о бок, что позволяет визуально сравнить финансовые результаты деятельности обоих ресторанов между собой и сделать предварительные предположения о том, в каком из них сумма чека является более вариабельной, в каком – менее. Данные первой выборки более «сжаты» вокруг медианы, а сумма чека – менее «разбросана», так как 50% всех заказов лежат в интервале

[400, 500] руб., в то время как в ресторане Б сумма чека в половине заказов, рассмотренных в выборке, находится в интервале [300, 600] руб., и этот интервал существенно шире. Диаграмма «Ящик с усами» – наглядная форма презентации данных, с помощью которой можно на одном рисунке отобразить до нескольких десятков диаграмм и сопоставить разные выборки друг с другом. Это создаёт колоссальные удобства в визуализации, «прочтении» и сравнительном анализе информации. ■

4.3 Задачи для тренировки

Задача 1. Пусть распределение оценок абитуриентов (по 10-балльной шкале), полученных ими на вступительных экзаменах по профильному предмету при поступлении, выглядит следующим образом:

Оценка, балл	0	1	2	3	4	5	6	7	8	9	10	Всего:
Число абитуриентов	80	140	160	180	180	200	540	300	140	60	20	2000

Таблица 4.1: Распределение оценок студентов

Найдите первый и девятый децили выборки.

Задача 2. Выборка представлена значениями

$$X : 10, 0, 12, 4, 5, 2, 2, 7, 13, 6, 7, 5, 6, 8, 15, 3.$$

Найдите медиану, нижнюю и верхнюю квартили, диапазон типичных значений. Укажите нетипичные значения в выборке, если они есть. Постройте диаграмму «Ящик с усами».

Задача 3 (Рост юношей и девушек). В ходе небольшого опроса были получены следующие результаты роста (с точностью до сантиметра):

Юноши: 169, 173, 181, 178, 172, 178, 178, 192;

Девушки: 172, 166, 174, 164, 176, 159, 161, 176, 175, 176, 167.

Для обеих выборок найдите медиану, нижнюю и верхнюю квартили, диапазон типичных значений. Укажите нетипичные значения в выборках, если они есть. Постройте диаграммы «Ящик с усами».

Задача 4. На представленных вам чертежах вы можете найти частотные диаграммы (рис. ??) и диаграммы «ящик с усами» (рис. ??). Соотнесите разные типы диаграмм, которые отображают один и тот же набор данных.

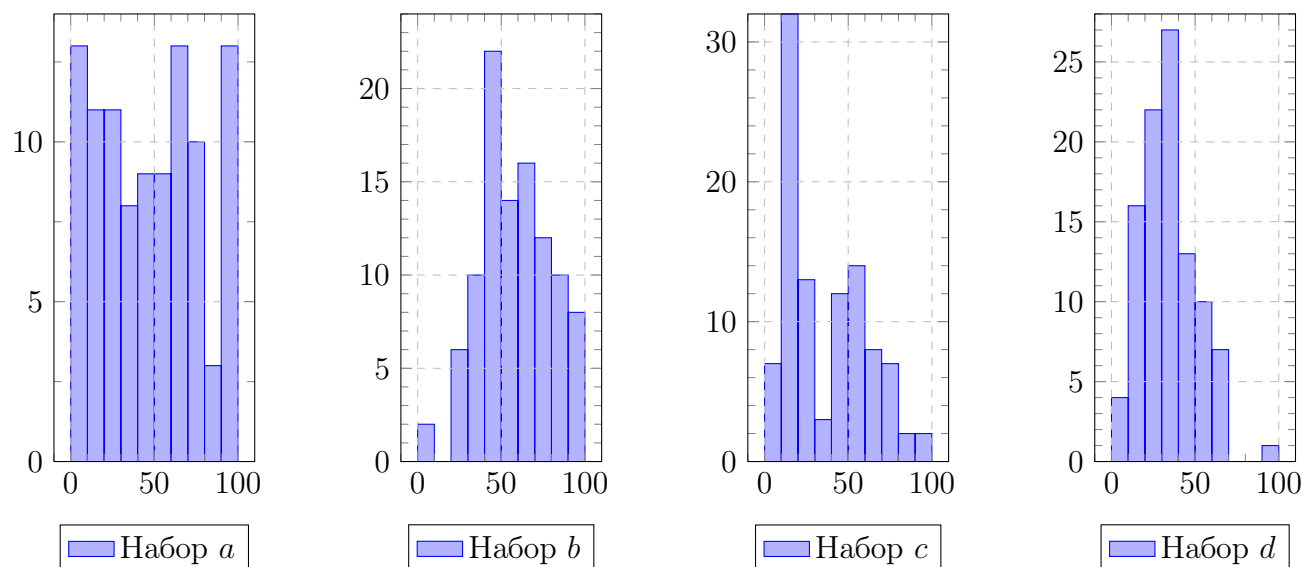


Рис. 4.1: Частотные диаграммы

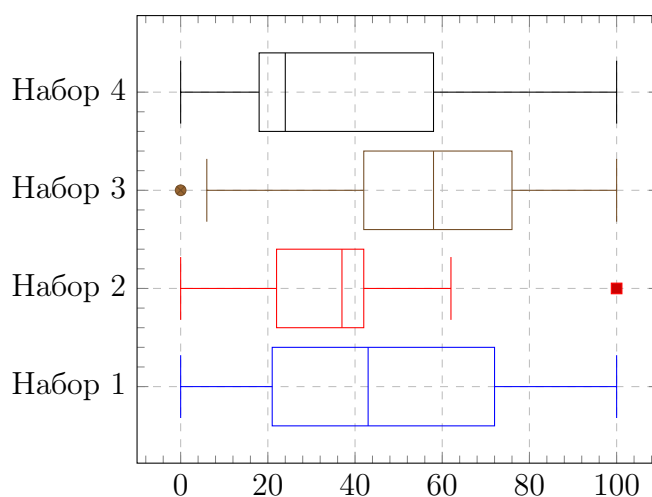


Рис. 4.2: Ящики с усами

4.4 Домашнее задание

Задача 1. Для следующих выборок найдите указанные числовые характеристики и постройте диаграммы «Ящик с усами». На основании диаграммы укажите: есть ли в выборке «подозрительные» наблюдения и/или выбросы.

1. Квартили;
2. Нижний и верхний дециль.

а) $A = (34, 48, 51, 52, 56, 59, 63, 64, 64, 66, 70, 70, 71, 74, 74, 75, 76, 80, 85, 85, 86, 88)$;

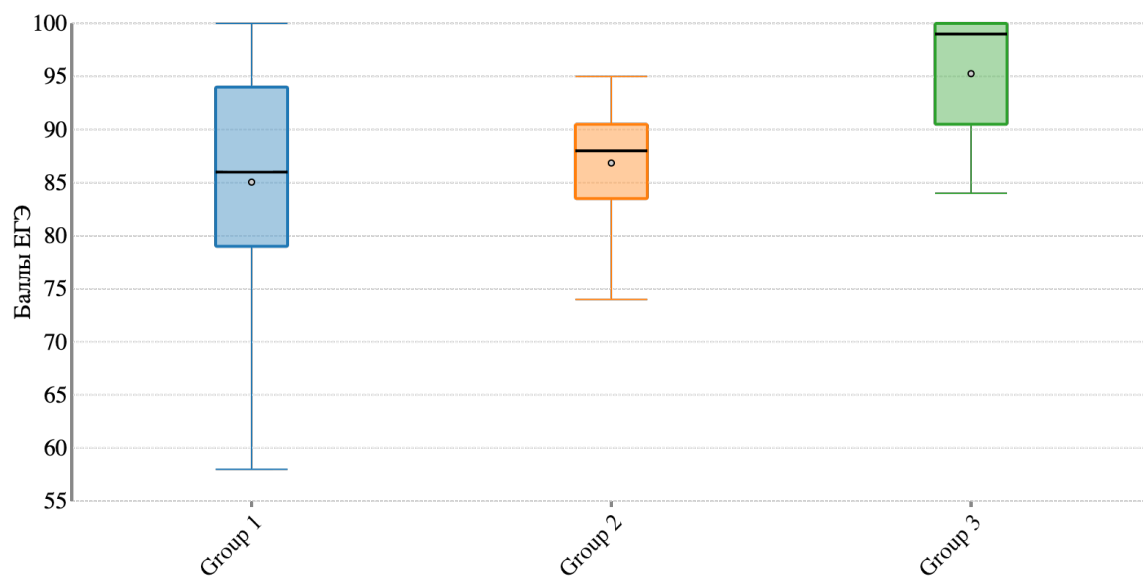
б) $B = (47, 50, 51, 55, 62, 62, 65, 66, 69, 71, 71, 72, 73, 75, 77, 80, 82, 82, 85, 89, 89, 90)$.

Задача 2. Выборки A_1 и A_2 – результаты ЕГЭ по профильной математике первой и второй групп 2023 года. Постройте диаграммы «Ящик с усами» и сделайте вывод.

$$A_1 = (100, 76, 80, 74, 70, 78, 70, 72, 80, 80, 78, 82, 82);$$

$$A_2 = (72, 72, 80, 68, 66, 78, 86, 70, 64, 68, 70, 90, 64, 70).$$

Задача 3. На рисунке ниже представлены ящики с усами баллов ЕГЭ по профильной математике для выпускников 2024 (Group 1), 2025 (Group 2) и 2026 года (Group 3). Сравните, если это возможно, дисперсии по выпускникам первых групп этих лет.



Тема 5

Основы статистики. Точечные оценки параметров

Статистика – это математическая теория того, как узнать нечто о мире через опыт.

В.Томпсон, английский статистик.

5.1 Теоретический материал

Предмет изучения статистики

Американский журнал «Литературное обозрение» провел опрос относительно исхода президентских выборов в США в 1936 году. Кандидатами на этих выборах были Ф.Д.Рузвельт и А.М.Ландон. В качестве генеральной совокупности редакция журнала использовала телефонные книги. Отобрав случайно 4 миллиона адресов, она разослала по всей стране открытки с вопросом об отношении к кандидатам в президенты. Затратив большую сумму на рассылку и обработку открыток, журнал объявил, что на предстоящих выборах президентом США с большим перевесом будет избран Ландон. Результат выборов оказался противоположным этому прогнозу. В чем они ошиблись?

Математическую статистику глобально интересуют два вопроса:

- Проверка статистических гипотез;
- Оценка некоторых параметров генеральной совокупности по параметрам выборки.

Например для первого пункта нам может быть интересно, есть ли зависимость между номером группы по математике и результатами независимой диагностики знаний обучающихся всех пяти групп, в предположении, что верна гипотеза

$$H_0 : \{ \text{чем меньше номер группы, тем выше балл за работу} \}.$$

Примером задачи второго типа служит понимание того, каким будет рейтинг кандидата на выборную должность после публикации некоторого пресс-релиза на основе полученного рейтинга от репрезентативной выборки фокус-группы.

Заметим, что чрезвычайно важно здесь слово *репрезентативность* – этим термином называют выборки, пропорционально представляющие важные для исследования категории генеральной совокупности. Например выборка, состоящая, в основном, из бабушек, сидящих на лавочках у подъезда для политического исследования нерепрезентативна.

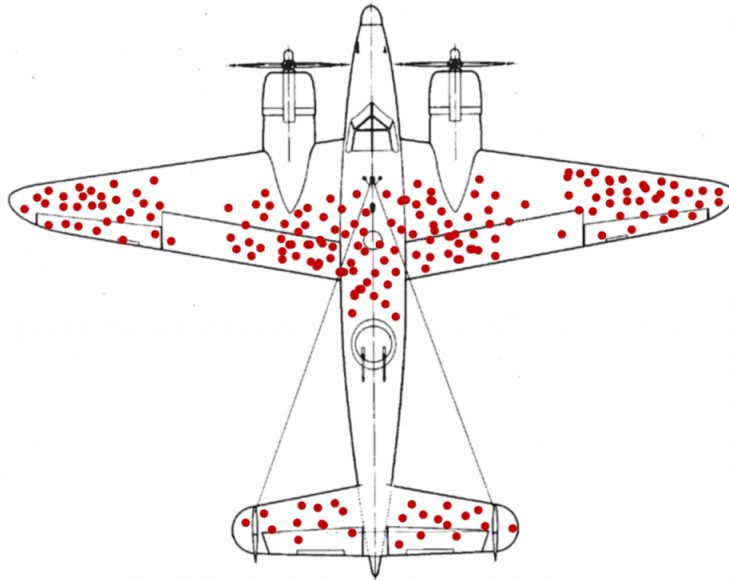


Рис. 5.1: Распределение пробоин на вернувшихся из боя самолётах

Нерепрезентативность выборки может приводить к так называемой ошибке выжившего. Во Вторую мировую войну американские военные столкнулись со следующей проблемой. Не все американские бомбардировщики после задания возвращались на базу. На вернувшихся самолетах оставалось множество пробоин от выстрелов противника (см. рис. ??), но распределены они были неравномерно: больше всего на фюзеляже и прочих частях, меньше в топливной системе и гораздо меньше – в двигателе.

Командованию казалось логичным, что в наиболее поврежденных местах нужно установить больше брони. Привлеченный к решению задачи математик Абрахам Вальд возразил: данные как раз показывают, что самолет, получивший пробоины в этих местах, еще может вернуться на базу. А самолет, которому попали в бензобак или двигатель, выходит из строя и не возвращается. Поэтому укреплять следует те места, которые у вернувшихся самолетов повреждены меньше всего. В данном примере, как вы уже догадались, из выборки напрасно исключены не вернувшиеся из боя самолёты.

Основные понятия

Определение 5.1: Выборка

Выборкой объема n при $n > 1$ называется случайный вектор

$$Z_n = (\xi_1, \dots, \xi_n),$$

компоненты которого ξ_i , $i \in \{1, \dots, n\}$, называемые элементами выборки, являются независимыми случайными величинами с одной и той же функцией распределения $F_\xi(x)$. Будем говорить, что выборка Z_n соответствует функции распределения $F_\xi(x)$.

Определение 5.2: Реализация выборки

Реализацией выборки называется неслучайный вектор $z_n = (x_1, \dots, x_n)$, компонентами которого являются реализации соответствующих элементов выборки ξ_i , $i \in \{1, \dots, n\}$.

Из определений выше вытекает, что реализацию выборки z_n можно также рассматривать как последовательность $x_1, \dots, x_n$ из n реализаций одной и той же случайной величины ξ , по-

лученных в серии из n независимых одинаковых опытов, проводимых в одинаковых условиях. Поэтому можно говорить, что выборка Z_n порождена наблюдаемой случайной величиной ξ , имеющей распределение $F_\xi(x) = F(x)$.

Определение 5.3: Параметр распределения

Параметром распределения θ случайной величины ξ называется любая числовая характеристика этой случайной величины (математическое ожидание, дисперсия и т. п.) или любая константа, явно входящая в выражение для функции распределения.

Определение 5.4: Статистика

Случайная величина $T = \varphi(Z_n)$, где $\varphi(z_n)$ – произвольная функция, определенная на выборочном пространстве, называется статистикой.

Определение 5.5: Точечная оценка

Точечной (выборочной) оценкой неизвестного параметра распределения Θ называется произвольная статистика $\hat{\Theta}(Z_n)$, построенная по выборке Z_n .

Определение 5.6: Несмещённая оценка

Оценка $\hat{\Theta}$ называется несмещённой оценкой параметра Θ , если математическое ожидание оценки равно оцениваемому параметру, то есть:

$$\mathbb{E}(\hat{\Theta}) = \Theta.$$

Определение 5.7: Состоятельная оценка

Оценка $\hat{\Theta}$ называется состоятельной оценкой параметра Θ , если она сходится по вероятности к оцениваемому параметру, то есть:

$$\mathbb{P} \left(\left| \hat{\Theta}_n - \Theta \right| < \varepsilon \right) \xrightarrow{n \rightarrow \infty} 1.$$

Оценка методом моментов

Определение 5.8: Моменты

k -м моментом случайной величины ξ является математическое ожидание этой случайной величины, возведенной в k -ую степень. Таким образом, $\mathbb{E}(\xi^k) = \mu_k$ – k -ый момент случайной величины ξ .

Теорема 5.1: Несмещённость и состоятельность выборочного среднего

Пусть имеется выборка из распределения с конечным первым моментом $\mathbb{E}(\xi_1) = \mu_1$. Тогда

1. $\mathbb{E}(\bar{X}) = \mathbb{E}(\xi)$, т. е. выборочное среднее $\bar{X}$ является несмещённой оценкой для истинного математического ожидания $\mathbb{E}(\xi)$;
2. $\bar{X} \xrightarrow{n \rightarrow \infty} \mathbb{E}(\xi)$ то есть выборочное среднее $\bar{X}$ является состоятельной оценкой для $\mathbb{E}(\xi)$.

Процедура оценивания методом моментов состоит в приравнивании k истинных моментов к K выборочным моментам для оценивания k неизвестных параметров модели $\Theta_1, \Theta_2, \dots, \Theta_k$.

Систему не всегда удаётся легко решить. И моменты можно брать разные, но мы договоримся брать моменты по порядку.

Оценка методом максимального правдоподобия

Пусть $Data = \{\xi_1, \xi_2, \dots, \xi_n\}$ – выборка из известного распределения вероятностей.

Правдоподобием выборки называется

$$(Data) = \mathbb{P}(\xi_1 \cap \xi_2 \cap \dots \cap \xi_n),$$

т.е. вероятность одновременно получить наблюдения (имеющуюся выборку). Поскольку выборка по определению – последовательность независимых одинаково распределённых случайных величин, то

$$\mathbb{P}(\xi_1 \cap \xi_2 \cap \dots \cap \xi_n) = \mathbb{P}(\xi = \xi_1) \cdot \mathbb{P}(\xi = \xi_2) \cdot \dots \cdot \mathbb{P}(\xi = \xi_n) = \prod_{i=1}^n \mathbb{P}(\xi = \xi_i).$$

Функция правдоподобия имеет большое практическое значение, поскольку на практике мы обычно не знаем параметров распределения (генеральной совокупности) и нуждаемся в их оценке. Функция правдоподобия позволяет построить такие оценки. А соответствующий метод называется методом максимального правдоподобия. Из названия метода следует, что требуется максимизировать функцию правдоподобия. То есть найти такое значение параметра, при котором функция правдоподобия будет максимальной.

Получающееся при этом значение параметра объявляется *оценкой максимального правдоподобия*.

Идея данного метода проистекает из представления о том, что выборка – это единственный источник знания о генеральной совокупности, доступный нам. Все, что мы знаем о генеральной совокупности, содержится в выборке. Поэтому нам ничего иного не остается, как сказать, что выборка – это наиболее точное отражение генеральной совокупности, доступное нам. А раз так, мы должны полностью положиться на выборку и понять, каков должен быть искомый параметр, чтобы появление именно такой выборки было наиболее вероятным. Итак, центральная идея ММП – это поиск такого значения параметра генеральной совокупности, при котором имеющаяся выборка становится наиболее вероятной.

Заметим, что часто для произведения степенных выражений искать максимум не представляется удобным, в таком случае на помощь приходит монотонная на области определения функция $f(x) = \ln x$ и логарифмическая функция правдоподобия $\ln(L(Data))$.

5.2 Примеры решения задач

Задача 1. Рассмотрим в роли оцениваемого параметра Θ дисперсию генеральной совокупности $\mathbb{D}(\xi)$. Пусть в качестве оценки параметра $\hat{\Theta}$ рассмотрена выборочная дисперсия. Является ли она несмещённой?

Решение.

Имеем

$$\hat{\Theta} = \frac{1}{n} \sum_{i=1}^n \xi_i^2 - \left(\frac{1}{n} \sum_{i=1}^n \xi_i \right)^2 = \frac{n-1}{n^2} \sum_{i=1}^n \xi_i^2 - \frac{2}{n^2} \sum_{i=1}^n \sum_{i < j} \xi_i \xi_j.$$

Перейдём к математическому ожиданию, воспользовавшись независимостью и одинаковым распределением всех ξ и тем, что $\mathbb{E}(\xi^2) = (\mathbb{E}(\xi))^2 + \mathbb{D}(\xi)$.

$$\begin{aligned}\mathbb{E}(\hat{\Theta}) &= \frac{n-1}{n^2} \mathbb{E} \left(\sum_{i=1}^n \xi_i^2 \right) - \frac{2}{n^2} \mathbb{E}(\xi_1 \xi_2 + \dots + \xi_{n-1} \xi_n) = \\ &= \frac{n-1}{n^2} \cdot n \mathbb{E}(\xi^2) - \frac{2}{n^2} \cdot \frac{n(n-1)}{2} \mathbb{E}(\xi)^2 = \frac{n-1}{n} ((\mathbb{E}\xi)^2 + \mathbb{D}(\xi)) - \frac{n-1}{n} (\mathbb{E}(\xi))^2 = \frac{n-1}{n} \cdot \mathbb{D}(\xi).\end{aligned}$$

Получили, что выборочная дисперсия – смещённая оценка истинной дисперсии. В качестве несмещённой окружности берут *исправленную* выборочную дисперсию:

$$\hat{\Theta} = \frac{n}{n-1} \mathbb{D}\xi.$$

Задача 2. Радиус окружности измеряется с погрешностью, которая распределена по нормальному закону со средним значением 0 и неизвестной дисперсией σ^2 . Найдите несмещённую оценку площади круга по выборке $X_1, \dots, X_n$ измерений радиусов.

Решение.

Задача 3. Пусть дана выборка из $B(10, 0.8)$, состоящая из следующих трех наблюдений: 5, 7, 9. То есть три раза независимым образом была проведена серия из 10 испытаний Бернулли с вероятностью успеха 0.8. Первая серия из 10 испытаний привела к 5 успехам (первое наблюдение в выборке); вторая – к 7 успехам; третья – к 10.

- Чему равно правдоподобие этой выборки?
- Методом максимального правдоподобия оцените параметр p распределения по данной выборке.

Решение.

- Пусть $\xi \sim B(10, 0.8)$. Тогда $\mathbb{P}(5, 7, 9) = C_{10}^5 0.8^5 0.2^5 \cdot C_{10}^7 0.8^7 0.2^3 \cdot C_{10}^9 0.8^9 0.2^1 \approx 0.001428$;
- Теперь выполним обратную процедуру. Пусть известно лишь что выборка получена из распределения $B(10, p)$. Требуется оценить параметр $\Theta = p$. Воспользуемся методом максимального правдоподобия. Имеем $L(5, 7, 9) = C_{10}^5 p^5 (1-p)^5 \cdot C_{10}^7 p^7 (1-p)^3 \cdot C_{10}^9 p^9 (1-p)$.

Для удобства рассмотрим логарифмическую функцию правдоподобия:

$$\begin{aligned}\ln(L(5, 7, 9)) &= \ln(C_{10}^5) + \ln(p^5) + \ln((1-p)^5) + \ln(C_{10}^7) + \ln(p^7) + \ln((1-p)^3) + \\ &+ \ln(C_{10}^9) + \ln(p^9) + \ln(1-p) = \ln(C_{10}^5) + \ln(C_{10}^7) + \ln(C_{10}^9) + 21 \ln(p) + 9 \ln(1-p).\end{aligned}$$

Для нахождения максимума найдём производную логарифмической функции правдоподобия:

$$(\ln(L(5, 7, 9)))' = \frac{21}{p} - \frac{9}{1-p}.$$

Приравнивая полученную производную к нулю, получаем значение $p = 0.7$. Проверкой убеждаемся в том, что $p = 0.7$ – точка максимума.

Хорошо или плохо удалось оценить истинный параметр? Нужно отметить, что *хорошо!* Ведь всего по трём реализациям случайной величины мы оценили истинный параметр с точностью 0.1.

Отметим также, что полученная оценка совпадает с частотной оценкой вероятности успеха в схеме испытаний Бернулли.

Задача 4. Предположим, у нас есть случайная выборка $X_1, X_2, \dots, X_n$, полученная из равномерного распределения на отрезке $[0, \Theta]$. Методом максимального правдоподобия найдите оценку параметра Θ .

Решение.

Плотность вероятности для каждого X_i задается как:

$$f(x_i; \Theta) = \begin{cases} \frac{1}{\Theta}, & \text{если } 0 \leq x_i \leq \Theta \\ 0, & \text{в противном случае} \end{cases}$$

Наша задача – оценить неизвестный параметр Θ , используя метод максимального правдоподобия (ММП).

Функция правдоподобия $L(\Theta; X_1, X_2, \dots, X_n)$ выражает вероятность наблюдения нашей выборки при заданном значении параметра Θ . Поскольку наблюдения независимы, правдоподобие равно произведению плотностей вероятности для каждого X_i :

$$L(\Theta; X_1, X_2, \dots, X_n) = f(X_1; \Theta) \cdot f(X_2; \Theta) \cdot \dots \cdot f(X_n; \Theta)$$

Подставляя нашу плотность вероятности равномерного распределения, получаем:

$$L(\Theta; X_1, X_2, \dots, X_n) = \begin{cases} \frac{1}{\Theta^n}, & \text{если } 0 \leq X_i \leq \Theta \text{ для всех } i \\ 0, & \text{в противном случае} \end{cases}$$

Чтобы правдоподобие было ненулевым, необходимо, чтобы все наблюдения X_i были не меньше 0 и не больше Θ . Это значит, что Θ должно быть не меньше, чем максимальное значение из нашей выборки.

Цель ММП – найти такое значение параметра Θ , которое максимизирует функцию правдоподобия $L(\Theta)$. В данном случае, $L(\Theta)$ равна $\frac{1}{\Theta^n}$ (или 0), и нам нужно максимизировать её по Θ . Поскольку Θ находится в знаменателе, максимизировать $L(\Theta)$ означает минимизировать Θ .

Однако, помним, что Θ должно быть не меньше, чем максимальное значение из нашей выборки:

$$\Theta \geq \max\{X_1, X_2, \dots, X_n\}$$

Оптимальным значением для Θ будет ровно максимальное значение нашей выборки. Любое меньшее значение сделает правдоподобие равным нулю, а любое большее – уменьшит правдоподобие.

Таким образом, оценка максимального правдоподобия для параметра Θ , обозначенная как $\hat{\Theta}$, равна:

$$\hat{\Theta} = \max\{X_1, X_2, \dots, X_n\}$$

5.3 Задачи для тренировки

Задача 1. Найдите несмещённую оценку для вероятности успеха в серии испытаний Бернулли.

Задача 2. По выборке x_1, x_2, x_3, x_4, x_5 из нормального распределения $\mathcal{N}(\theta, \sigma^2)$ построена следующая оценка параметра θ : $\hat{\theta} = 0.1x_1 + 0.2x_2 + 0.3x_3 + 0.3x_4 + 0.1x_5$. Выяснить является ли оценка $\hat{\theta}$ несмещённой.

Задача 3. По выборке (см. таб. ??) из равномерного распределения $U_{[a,58]}$ найдите несмещённую оценку для параметра $\Theta = a$.

41.8 | 43.2 | 49.3 | 21.8 | 19.1 | 7.4 | 6.5 | 25.7

Таблица 5.1: Случайная выборка из распределения $U_{[a,58]}$

Задача 4. С помощью выборочного исследования изучают среднее время стоянки автомобиля на паркинге торгового центра. Получена следующая случайная выборка:

56, 131, 83, 32, 77 и 110 мин.

Сделайте оценку стандартного отклонения времени стоянки автомобиля на паркинге на основе несмещённой оценки дисперсии. Результат округлите до целого числа минут.

Задача 5. Число семян сорняков в пробах зерна подчинено закону Пуассона. Имеется выборка $N = 130$ проб зерна. Результаты записаны в таблице ?. Найти несмещённую оценку параметра λ по выборке.

ξ_i	0	1	2	3	4	5
n_i	9	39	40	24	11	7

Таблица 5.2: Выборка проб зерна

Задача 6. Пусть выборка $\xi_1, \xi_2, \dots, \xi_n$ – реализации геометрически распределённой случайной величины $\xi \sim \text{Geom}(p)$. Методом максимального правдоподобия найдите точечную оценку параметра p .

Задача 7. Случайная величина X распределена по закону Пуассона. Результаты 464 независимых наблюдений X отражены в таблице:

Значение X	0	1	2	3
Частота n_i	205	155	78	26

Найдите методом моментов точечную оценку $\hat{\lambda}$ параметра λ .

Задача 8. Случайная величина X распределена по геометрическому закону. Результаты 15 независимых наблюдений X отражены в таблице

Значение X	2	3	4	5
Частота n_i	1	6	5	3

Найдите методом моментов точечную оценку $\hat{p}$ вероятности p успеха в одном испытании.

Задача 9. Случайная величина X распределена по биномиальному закону. Результаты наблюдений X в 100 опытах, в каждом из которых было проведено 10 испытаний, отражены в таблице

Значение X	0	1	2	3	4	5	6	7
Частота n_i	1	5	10	20	25	10	5	10

Найдите методом моментов точечную оценку $\hat{p}$ вероятности p успеха в одном испытании.

Задача 10. Случайная величина X распределена по показательному закону. Результаты наблюдений X в 1000 опытах отражены в таблице

Значение X	5	15	25	35	45	55	65
Частота n_i	370	240	155	95	75	40	25

Найдите методом моментов точечную оценку $\hat{\lambda}$ параметра λ .

Задача 11. Используя метод моментов, оцените параметры a и b равномерного распределения на отрезке по эмпирическому распределению

Значение X	2	3	4	5	6
Частота n_i	4	6	5	12	8

5.4 Домашнее задание

Задача 1. Колл-центр банка принимает в среднем λ звонков в час. Какова точечная оценка параметра $\Theta = \lambda$ распределения Пуассона, описывающего количество звонков в час, если за 2 часа было принято 18 звонков?

Задача 2. Баскетболист бросает мяч в корзину до первого попадания. Вероятность попадания при каждом броске постоянна. Известно, что первое попадание произошло с 4-го броска. Оцените вероятность попадания в корзину при каждом броске, используя оценку максимального правдоподобия и оценку, полученную методом моментов.

Задача 3. Пусть измерения $X_1, X_2, \dots, X_n$ независимы и каждое имеет нормальное распределение с математическим ожиданием R (истинный радиус шара) и известной дисперсией σ^2 , т. е.

$$X_i \sim \mathcal{N}(R, \sigma^2).$$

Требуется найти несмещённую оценку для объёма шара

$$V = \frac{4}{3}\pi R^3.$$

Задача 4. Случайная величина X распределена по закону Пуассона. Результаты 468 независимых наблюдений X отражены в таблице

Значение X	0	1	2	3
Частота n_i	205	160	81	22

Найдите методом моментов точечную оценку $\hat{\lambda}$ параметра λ .

Задача 5. Случайная величина X распределена по геометрическому закону. Результаты 10 независимых наблюдений X отражены в таблице

Значение X	1	2	3	4
Частота n_i	1	4	3	2

Найдите методом моментов точечную оценку $\hat{p}$ вероятности p успеха в одном испытании.

Задача 6. Случайная величина X распределена по биномиальному закону. Результаты наблюдений X в 100 опытах, в каждом из которых было проведено 10 испытаний, отражены в таблице

Значение X	0	1	2	3	4	5	6	7
Частота n_i	1	5	20	10	5	10	25	10

Найдите методом моментов точечную оценку $\hat{p}$ вероятности p успеха в одном испытании.

Задача 7. Случайная величина X распределена по показательному закону. Результаты наблюдений X в 1000 опытах отражены в таблице

Значение X	5	15	25	35	45	55	65
Частота n_i	380	230	160	90	70	40	30

Найдите методом моментов точечную оценку $\hat{\lambda}$ параметра λ .

Задача 8. Используя метод моментов, оцените параметры a и b равномерного распределения на отрезке по эмпирическому распределению

Значение X	3	5	7	9
Частота n_i	21	18	15	26

Задача 9. Случайная величина X подчиняется нормальному закону распределения. Результаты наблюдений X отражены в таблице

Значение X	0	1	2	3	4	5	6	7	8	9	10
Частота n_i	8	11	13	23	42	59	48	27	16	14	7

Используя метод моментов, оценить параметр m , а также получить смещённую $\hat{\sigma}$ и несмещённую $\hat{s}$ оценки параметра σ .

Тема 6

Линейная регрессия

6.1 Теоретические материалы

Диаграммы рассеивания

Определение 6.1: Диаграмма рассеивания

Диаграмма рассеивания – способ представления двух совместно наблюдаемых величин точками на координатной плоскости. Когда точек много, они образуют облако рассеивания.

Часто бывает полезно знать, есть ли некоторая связь между изучаемыми величинами и, если есть, то какова она. Разобраться в этом помогает диаграмма рассеивания. Покажем на примере, как её построить.

Есть ли связь между ростом и массой человека? Для наглядного ответа построим диаграмму рассеивания. Данными для этой диаграммы служат пары величин. Каждая пара – это рост и масса одного человека. В таблице ?? даны значения роста и массы 15 юношей.

Рост, см	167	169	179	178	177	175	171	181
Вес, кг	62	67	70	72	70	69	63	80
Рост, см	174	175	180	174	172	178	171	
Вес, кг	73	66	75	70	67	74	66	

Таблица 6.1: Роста и масса 15 юношей.

Чтобы получить диаграмму рассеивания, нужно в системе координат построить точки, абсциссы которых – рост, а ординаты – соответствующая масса. Получаем диаграмму ??.

Точки на диаграмме образуют облако рассеивания, или, более поэтично, созвездие. Видно, что на построенной диаграмме точки группируются вдоль некоторой наклонной прямой, направленной вправо и вверх.

Это значит, что между величинами «рост» и «масса», скорее всего, имеется положительная связь: в целом чем человек выше, тем больше его масса. Разумеется, в каждом конкретном случае это правило может нарушаться. Например, в таблице есть юноша ростом 179 см и массой 70 кг и юноша ростом 174 см и массой 73 кг. Но такая ситуация скорее исключение.

При этом важно помнить, что диаграмма рассеивания – не доказательство наличия или отсутствия связи. С помощью диаграммы мы лишь можем предположить, что связь есть или что её нет. Чем больше наблюдений, тем более обоснованно предположение. Наличие связи не означает, что одна величина обусловлена другой. Наблюдаемая связь далеко не всегда причинно-

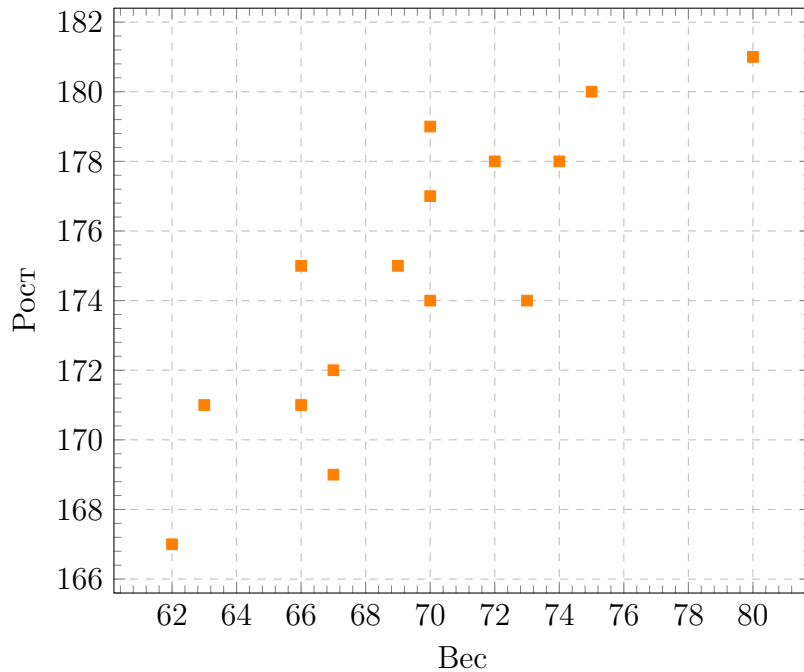


Рис. 6.1: Диаграмма Рост/Вес

следственная. Механизм связи может быть сложным. Часто обе величины зависят не друг от друга, а от каких-либо ещё факторов, которые не входят в исследование.

Линейная регрессия и метод наименьших квадратов

Предположим, у нас есть серия данных, полученных в результате эксперимента, которые измеряют положение (x) в зависимости от времени (t) частицы, движущейся с постоянной скоростью.

В диаграмме ?? интуитивно мы видим, что между ними есть связь – все они визуально складываются в линию. И тут возникает хитрый вопрос: «А как это доказать?»

Решение простое:

- провести несколько линий;
- придумать какой-то критерий для сравнения;
- сравнить их между собой и выбрать ту, которая будет максимально близка к этому критерию.

Эта линия и есть линейная регрессия – описание соответствия между случайными переменными. (см. рис. ??)

Решим общую задачу.

Пусть нам даны точки на плоскости $\{(x_1, y_1), \dots, (x_n, y_n)\}$ и мы ищем такую линейную функцию $y = a + b \cdot x$, чтобы ее график ближе всего находился к точкам.

Как видно из иллюстрации ??, слово «ближе», то есть расстояние от точки до прямой можно понимать по-разному, например геометрически – это длина перпендикуляра. Однако в контексте нашей задачи нам нужно *функциональное* расстояние, а не геометрическое.

Нас интересует разница между экспериментальным значением и предсказанием модели для каждого x_i , поэтому измерять нужно вдоль оси y .

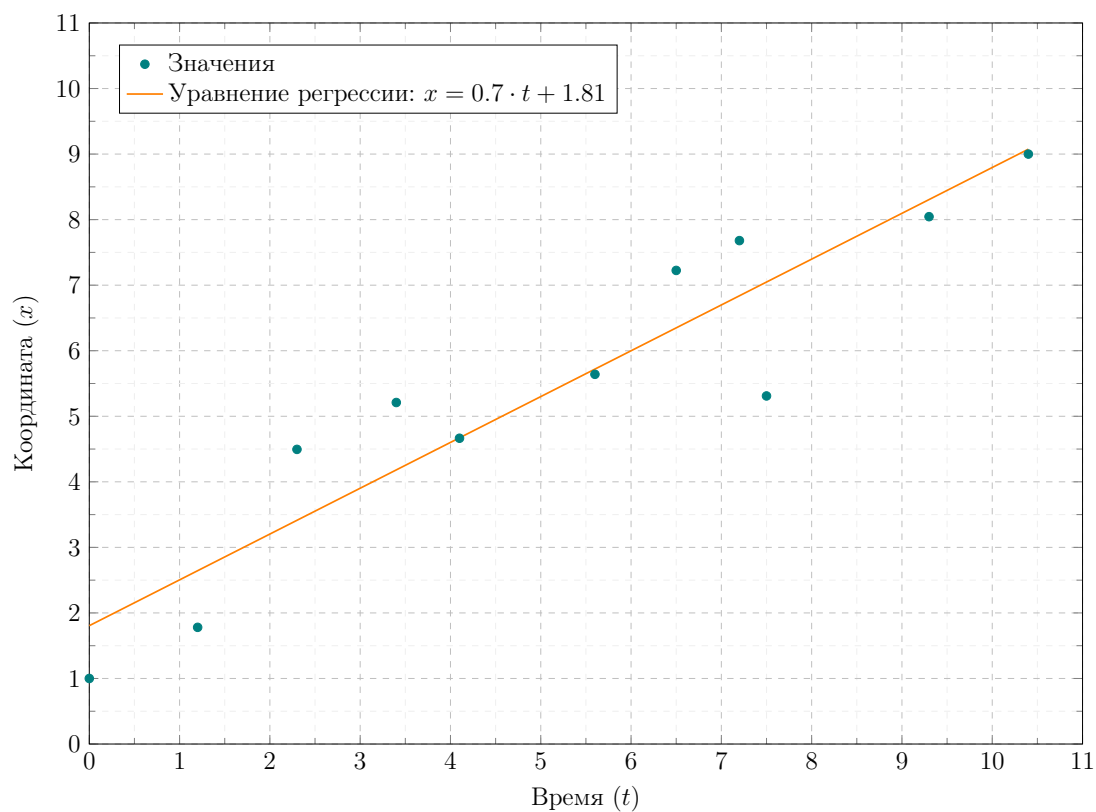


Рис. 6.2: Линейная регрессия

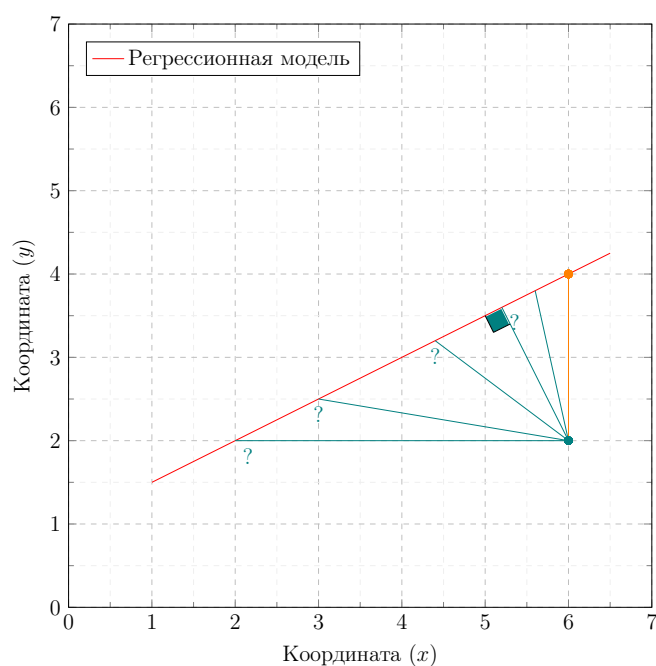


Рис. 6.3: Выбор расстояния

Первое, что приходит в голову, в качестве функции потерь попробовать выражение, зависящее от абсолютных значений разниц $|f(x_i) - y_i|$. Простейший вариант – сумма модулей отклонений $\sum_i |f(x_i) - y_i|$ приводит к Least Absolute Distance (LAD) регрессии.

Впрочем, более популярная функция потерь – сумма квадратов отклонений регрессанта от

модели. В англоязычной литературе она носит название Sum of Squared Errors (SSE)

$$\text{SSE}(a, b) = \text{SS}_{\text{residuals}} = \sum_{i=1}^N \text{отклонение}_i^2 = \sum_{i=1}^N (y_i - f(x_i))^2 = \sum_{i=1}^N (y_i - a - b \cdot x_i)^2,$$

Метод наименьших квадратов (по англ. OLS) – линейная регрессия с $\text{SSE}(a, b)$ в качестве функции потерь.

Как же минимизировать сумму отклонений? Оказывается, можно минимизировать её и явно найти значения a и b .

Рассмотрим всевозможные функции $L(a, b) = \sum_{i=1}^N (y_i - a - b \cdot x_i)^2$ для фиксированных выборок X и Y и выберем такую, которая на данных выборках принимает наименьшее значение. Для этого возьмём частные производные и приравняем их к нулю:

$$\begin{cases} \frac{\partial L}{\partial a} = - \sum_{i=1}^N 2(y_i - a - b \cdot x_i), \\ \frac{\partial L}{\partial b} = - \sum_{i=1}^N 2x_i(y_i - a - b \cdot x_i); \end{cases} \implies \begin{cases} \bar{y} - a - b\bar{x} = 0, \\ \overline{xy} - a\bar{x} - b\bar{x}^2 = 0; \end{cases} \iff \begin{cases} \bar{y} = a + b\bar{x}, \\ \overline{xy} = a\bar{x} + b\bar{x}^2. \end{cases}$$

Домножим первое уравнение системы на $\bar{x}$ и вычтем второе уравнение из первого, получим:

$$\bar{x} \cdot \bar{y} - \overline{xy} = b(\bar{x}^2 - \overline{x^2}) \iff b = \frac{\text{cov}(x, y)}{\mathbb{D}(x)}.$$

Получили:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\mathbb{D}(x)}, \quad a = \bar{y} - b \cdot \bar{x}.$$

Тот же результат можно получить проще:

Теорема 6.1: SSE модель регрессии

Доказательство. Пусть имеются наблюдения (x_k, y_k) , $k = 1, \dots, n$. Ищем прямую $y = a + bx$, минимизирующую сумму квадратов отклонений:

$$S = \sum_{k=1}^n \varepsilon_k^2, \quad \varepsilon_k = y_k - (bx_k + a).$$

Сумма квадратов чисел ε_k наименьшая тогда, когда наименьшим является средний квадрат

$$\overline{\varepsilon^2} = \frac{1}{n} \sum_{k=1}^n \varepsilon_k^2.$$

Средний квадрат связан с дисперсией:

$$\overline{\varepsilon^2} = S_e^2 + \bar{\varepsilon}^2,$$

где $\bar{\varepsilon}$ — среднее арифметическое отклонений, а S_e^2 — их выборочная дисперсия.

При сдвиге прямой вверх или вниз все отклонения ε_k изменяются на одну и ту же величину, поэтому их дисперсия S_e^2 не меняется. Следовательно, $\overline{\varepsilon^2}$ минимально, когда $\bar{\varepsilon}^2$ минимально, т.е. $\bar{\varepsilon} = 0$. Отсюда

$$\bar{y} - b\bar{x} - a = 0 \implies a = \bar{y} - b\bar{x}.$$

Теперь подставим $a = \bar{y} - b\bar{x}$ в выражение для дисперсии отклонений. Заметим, что

$$S_e^2 = \frac{1}{n} \sum_{k=1}^n (y_k - bx_k - a)^2 = \frac{1}{n} \sum_{k=1}^n ((y_k - bx_k) - (\bar{y} - b\bar{x}))^2 = \frac{1}{n} \sum_{k=1}^n ((y_k - \bar{y}) - b(x_k - \bar{x}))^2.$$

Раскрывая квадрат, получаем

$$S_e^2 = S_y^2 - 2b \operatorname{cov}(x, y) + b^2 S_x^2,$$

где

$$S_x^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2, \quad S_y^2 = \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2, \quad \operatorname{cov}(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}).$$

Это квадратичная функция от b с положительным коэффициентом при b^2 , поэтому её минимум достигается при

$$b = \frac{\operatorname{cov}(x, y)}{S_x^2}.$$

Тогда

$$a = \bar{y} - b\bar{x}.$$

Таким образом, выборочные оценки коэффициентов регрессии имеют вид

$$\hat{b} = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Таким образом введём следующее определение:

Определение 6.2: Уравнение линейной регрессии

Уравнение линейной регрессии X на Y записывается в виде

$$x = \bar{x} + \rho(X, Y) \cdot \frac{\sigma(X)}{\sigma(Y)}(y - \bar{y}).$$

При желании вычислить регрессию Y на X , достаточно заменить X на Y и наоборот.

Важность критического мышления

При работе с таким мощным инструментом как линейная регрессия важно критически относиться к тому, что вы собираетесь ему «скормить». Хорошим примером тому является *Квартет Энскомба* (см. рис. ??) – четыре набора числовых данных, у которых простые статистические свойства идентичны, но их графики существенно отличаются.

Важно помнить, что алгоритм головы не имеет и вычисляя линейную регрессию для набора данных мы невольно можем найти связь там, где её нет.

6.2 Примеры решения задач

Задача 1. Известно, что между параметрами X и Y имеется зависимость, близкая к линейной. Пользуясь методом линейной регрессии, найдите линейное приближение зависимости $y(x)$ и на основании найденной зависимости сделайте вывод о характере зависимости: прямой или обратной. Выборка:

x	10	12	13.5	17	12.5	15	16	11	21	20
y	6	8	9	11	8	10	12	5	15	14

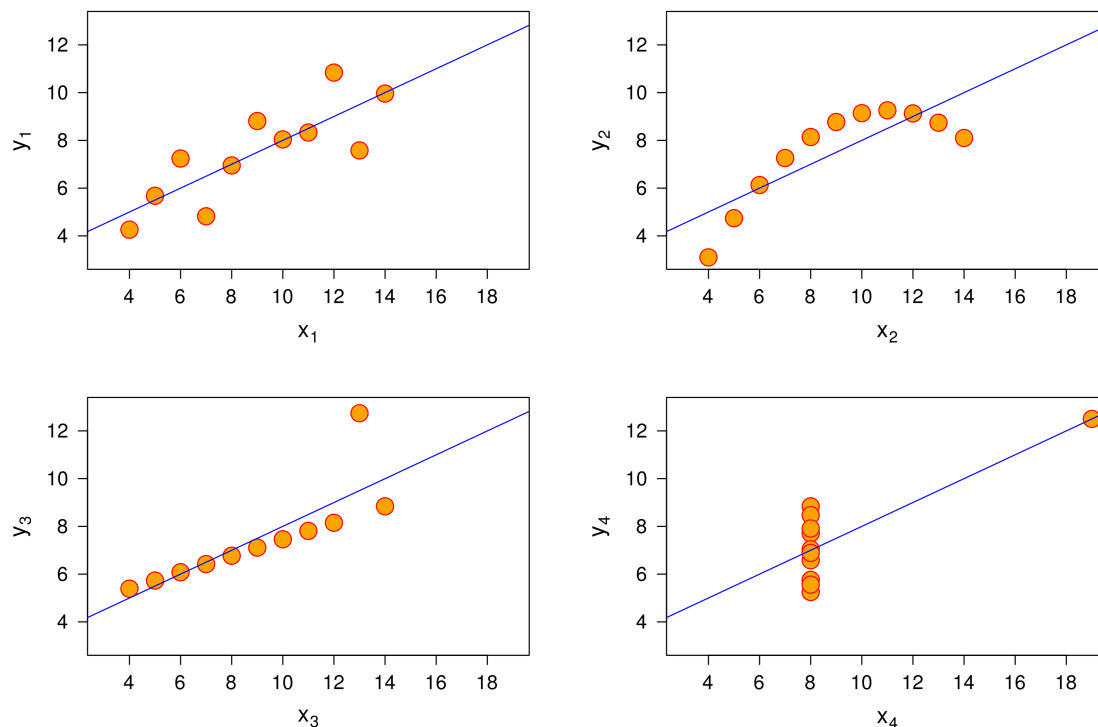


Рис. 6.4: Квартет Энского

Решение.

Оформим все необходимые вычисления в виде таблицы:

x	y	x^2	xy
10	6	100	60
12	8	144	96
13.5	9	182.25	121.5
17	11	289	187
12.5	8	156.25	100
15	10	225	150
16	12	256	192
11	5	121	55
21	15	441	315
20	14	400	280

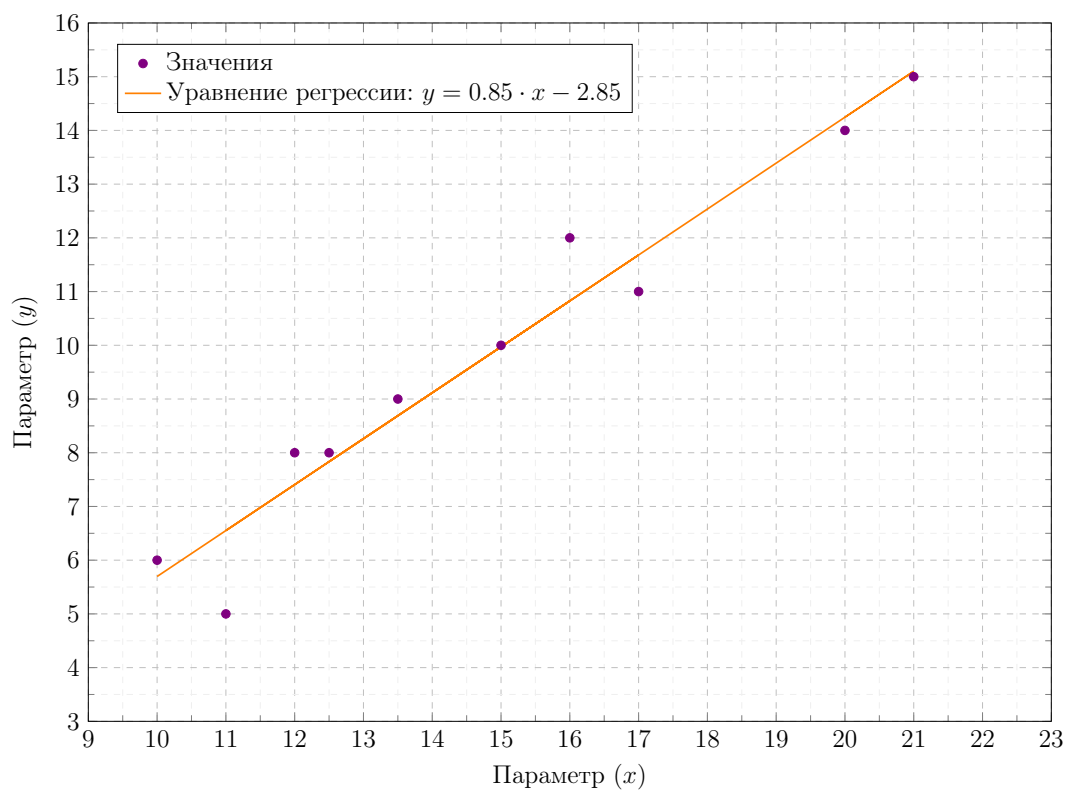
Остаётся вычислить $\bar{x} = 14.8$, $\bar{y} = 9.8$, $\overline{xy} = 155.65$ и $\overline{x^2} = 231.45$.

По формулам находим коэффициенты:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{155.65 - 14.8 \cdot 9.8}{231.45 - (14.8)^2} = \frac{1061}{1241}, \quad a = \bar{y} - b \cdot \bar{x} = 9.8 - \frac{1061}{1241} \cdot 14.8 = -\frac{3541}{1241}.$$

Для приближённого вычисления округлим найденные коэффициенты до сотых: $a = -2.85$, $b = 0.85$. Найденную прямую можно увидеть на рисунке ??.

Заметим, что обнаруженная зависимость прямая.



6.3 Задачи для тренировки

Задача 1. Покажите, что график линейной регрессии $y = a + bx$ проходит через точку $(\bar{x}, \bar{y})$.

6.4 Домашнее задание

Тема 7

Теорема Байеса в задачах

Меньше знаешь – крепче спишь

Поговорка

7.1 Теоретический материал

Теорема 7.1: Формула полной вероятности

Пусть все вероятностное пространство Ω разбито на конечный набор попарно непересекающихся событий A_i . Тогда для любого случайного события B :

$$\mathbb{P}(B) = \mathbb{P}(A_1) \cdot \mathbb{P}(B|A_1) + \mathbb{P}(A_2) \cdot \mathbb{P}(B|A_2) + \dots + \mathbb{P}(A_n) \cdot \mathbb{P}(B|A_n).$$

Иллюстрация этой формулы на рисунке ниже.

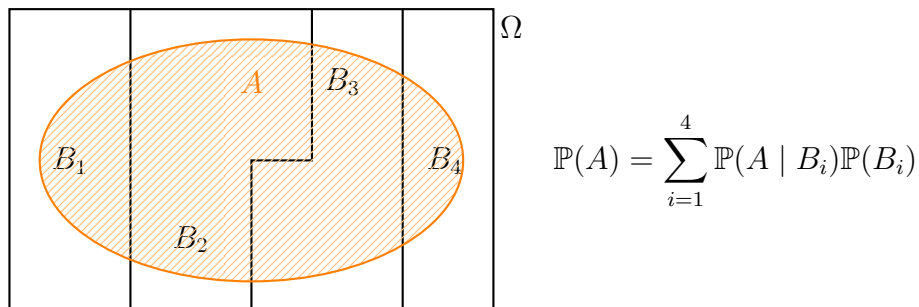


Рис. 7.1: Формула полной вероятности схематично

Теорема 7.2: Байес

Из представления $\mathbb{P}(A) \cdot \mathbb{P}(B|A) = \mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot \mathbb{P}(A|B)$ следует теорема Байеса:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

Формула Байеса позволяет переоценить вероятность событий уже после того, как эксперимент проведен. Она иллюстрирует сужение вероятностного пространства.

В качестве элементарного примера к теореме Байеса рассмотрим следующий: пусть на полке стоит много-много, например сотня, книг, и ровно в одной из них вы спрятали записку на

чёрный день – 100 рублей. Спрятали вы давно, опознать книгу возможности нет, и потому будем считать, что ваша значка в одной случайной книги с полки, причём в каждой книге она может быть равновероятно. Значит найти нужную книгу с первой попытки можно с вероятностью $p_1 = \frac{1}{100}$. Довольно малая вероятность и, скажем прямо, скорее всего вам не повезло. Но вы полны решимости найти заветную сотню и потому отбрасываете пустую книгу в сторону, узнав, что в ней значки не оказалось. В этот момент и происходит сужение вероятностного пространства: с значкой книга, по-прежнему.

7.2 Задачи для тренировки

Задача 1. У Буратино в правом кармане 4 серебряных и 2 золотые монеты. Буратино, не глядя, перекладывает три каких-то монеты в левый карман. Какова вероятность того, что обе золотые монеты оказались в одном кармане?

Задача 2. Самостоятельная работа состоит из двух тестовых вопросов. В каждом вопросе 5 вариантов ответа, из которых один – верный. Вовочка плохо готовился и поэтому знает только 10 вопросов (из 25). Он вынимает два вопроса случайно и без возвращения. Неправильный ответ на тесте не штрафуются, поэтому если Вовочка не знает ответа, то он отвечает наугад равновероятно. Какова вероятность того, что Вовочка знает ответ хотя бы на один вопрос, если он отметил верные ответы на оба из них?

Задача 3. Можно считать вероятности рождения девочек и мальчиков равными. В семьях Ивановых, Петровых и Сидоровых по два ребёнка. Про семью Ивановых известно, что старший ребёнок – девочка. Про семью Петровых – что хоть один ребёнок в семье – девочка. Наконец про семью Сидоровых известно, что хотя бы один из детей в семье – девочка, рождённая зимой. Найдите вероятности того, что в семьях Ивановых, Петровых и Сидоровых оба ребёнка – девочки.

Задача 4. Ниже приведены задача и её «решение». Оцените решение и, если оно содержит ошибки, исправьте их.

Задача. Первый стрелок попадает с вероятностью 0.2, второй – с вероятностью 0.6. Известно, что попал ровно один. Какова вероятность, что промахнулся первый?

«Решение» Первый промахивается с вероятностью 0.8, второй – с вероятностью 0.4. Значит, первый промахивается вдвое чаще. При условии, что промахнулся ровно один, вероятность промаха первого равна $\frac{2}{3}$.

Задача 5. Ниже приведены задача и её «решение». Оцените решение и, если оно содержит ошибки, исправьте их.

Задача Есть три карточки: красно-красная, сине-синяя и красно-синяя. Наугад берут одну карточку и наугад показывают одну из её сторон. Показанная сторона красная. Какова вероятность, что другая сторона тоже красная?

«Решение» Сине-синяя карточка отпадает. Остаются красно-красная и красно-синяя карточки; они были выбраны равновероятно. Только в первом случае другая сторона красная. Поэтому ответ равен $\frac{1}{2}$.

Задача 6. Пусть G – событие, при котором определенное лицо виновно в определенном ограблении. При сборе доказательств выясняется, что произошло событие E_1 , а чуть позже также выясняется, что произошло другое событие E_2 . Возможно ли, что по отдельности эти улики увеличивают вероятность вины (то есть $\mathbb{P}(G|E_1) > \mathbb{P}(G)$ и $\mathbb{P}(G|E_2) > \mathbb{P}(G)$), но вместе они уменьшают вероятность вины: $\mathbb{P}(G|E_1, E_2) < \mathbb{P}(G)$?

Задача 7. Объясните следующее высказывание Шерлока Холмса в терминах условной вероятности, тщательно проводя различие между предшествующей и последующей вероятностями: «Если исключить невозможное, то, что останется, и будет правдой, сколь бы невероятным оно ни казалось».

Задача 8. Покажите, что условие $\mathbb{P}(A|B) \leq \mathbb{P}(A)$ влечёт $\mathbb{P}(A|\overline{B}) \geq \mathbb{P}(A)$. Дайте интуитивную интерпретацию этому факту.

Задача 9. Верно ли, что для независимых событий A_1, A_2 и произвольного события B выполнено

$$P(A_1 \cap A_2 | B) = P(A_1 | B)P(A_2 | B)?$$

Задача 10. При подозрении на наличие некоторого заболевания пациента отправляют на ПЦР-тест. Если заболевание есть, то тест подтверждает его в 86% случаев. Если нет, то тест выявляет отсутствие заболевания в среднем в 94% случаев.

Известно, что в среднем тест оказывается положительным у 10% пациентов, направленных на тестирование. При обследовании некоторого пациента врач направил его на ПЦР-тест, который оказался положительным. Какова вероятность того, что пациент действительно имеет это заболевание?

Задача 11. Энтомолог предполагает, что найденная им бабочка может относиться к редкому подвиду бабочек, так как узор на крыльях у неё несимметричен. Известно, что в редком подвиде у 99% бабочек узор на крыльях несимметричен, а среди обычных бабочек только 0.5% имеют несимметричный узор. Также известно, что бабочек редкого подвида насчитывается лишь 0.2% среди всей популяции. Какова вероятность того, что найденная энтомологом бабочка относится к редкому виду?

Задача 12. Известно, что в результате десяти бросаний игральной кости выпала по крайней мере одна «шестерка». Какова вероятность того, что число выпавших «шестерок» больше единицы?

Задача 13. В первом загоне находится 10 овец, из которых 2 белые, остальные – черные, а во втором загоне – 7 белых и 3 черных овцы. Между загонами есть переход, и одна из овец первого загона перешла во второй. Переход закрыли и из второго загона отобрали одну овцу для стрижки. Найдите вероятность того, что пойманная овца оказалась из первого загона, если известно, что она белая?

Задача 14 (XVII Заочный конкурс учителей математики). Перед Атосом стоят шесть бокалов вина, которые он осушает один за другим. В четыре бокала Миледи подсыпала яд. Атос умрёт, если выпьет два отравленных бокала. Он уже выпил два бокала и пока жив. Каковы его шансы выжить, если он выпьет ещё один бокал?

Задача 15. В понедельник у Миши пять уроков, а во вторник – шесть. Чтобы определить, на каких уроках непредсказуемые учителя спросят у него домашнее задание, Миша бросает монету 11 раз – по числу возможных неприятностей. Если выпадает орёл, Миша считает, что на этом уроке его спросят, если решка – то не спросят. После уроков во вторник Миша констатировал, что ему удалось угадать 7 раз. Найдите вероятность того, что ровно 3 верных предсказания относились к понедельнику.

Задача 16 (Голосовой помощник). Распознавание речи часто сталкивается с проблемой омофонии. Для её решения применяется система, основанная на формуле Байеса. Так, например, во французском языке слова "mer", "m'ere" и "maire" (обозначим эти слова W_1 , W_2 и W_3 соответственно) разными носителями произносятся чуть по-разному, но в целом очень похоже.

Для обучения системы записывают различные варианты произношения, и по полученным данным оценивают вероятности для разных слов получить конкретное произношение A . Дополнительно, при распознавании не одного слова, а хотя бы фразы, из контекста можно оценить вероятности появления каждого из слов (или же считать их равновероятными, если оценка по контексту невозможна). Пусть для слов выше вариант произношения A встречается у 40% носителей для слова "mer", у 60% носителей для слова "m'ere" и у 50% носителей для слова "maire". Найти вероятности, что было произнесено каждое из слов при условии получения варианта произношения A и выбрать из них предпочтительное, если:

- а) нет дополнительной информации о контексте;
- б) анализ контекста даёт следующие результаты: наиболее вероятным является слово "mer" (40%), а самая низкая вероятность встретиться у слова "m'ere" (26%).

Задача 17 (Бунт ушастых). Ровно половина населения острова Невезения – зайцы, а все остальные – кролики. Если житель острова Невезения что-нибудь утверждает, он всегда искренне верит в то, что говорит. При этом зайцы добросовестно заблуждаются в среднем в каждом четвертом случае, а кролики добросовестно заблуждаются в среднем в каждом третьем случае. Однажды в центр острова вышел зверь и закричал: «Я не заяц!». Подумал и грустно сказал: «Я не кролик». Какова вероятность того, что он все же заяц?

Задача 18 (Кролики из олимпиады по теории вероятностей 2017). Кролики бывают гладкошерстные и мохнатые. Тип шерсти определяется генами, унаследованными от родителей. Каждый кролик имеет два аллеля одного и того же гена, отвечающего за тип шерсти. Аллель мохнатости обозначим H (или h), а аллель гладкошерстности обозначим S (или s). Если у кролика разные аллели, то все зависит от того, какой из аллелей доминантный, а какой – рецессивный. Каждый из родителей передает детенышу один из своих аллелей. Например, если отец имеет генотип hh , а мать – hS , то от отца детеныш получает h , а от матери равновероятно S или h . Известно, что в некоторой популяции кроликов аллель мохнатости встречается с вероятностью $p = 0.1$. Однажды в этой популяции при скрещивании мохнатого кролика с гладкошерстной крольчихой все четыре детеныша оказались мохматыми. Какой генотип наиболее вероятен у родителей этих крольчат?

Задача 19. Группа, в которой 27 студентов, сдает письменный зачет по решению олимпиадных задач. Преподаватель заготовил зачетную работу в 3 вариантах и распределяет варианты случайным образом с единственным условием: количество тех, кто получил разные варианты, должно отличаться не больше чем на единицу. Перед зачетом между студентами Сергеем и Алексеем состоялся следующий диалог.

С: Вероятность того, что у нас с тобой окажется один и тот же вариант, равна $1/3$.

А: Если сегодня кто-то один не придёт на зачет, то она окажется меньше, чем была бы, если бы пришли все.

С: Если сегодня не придут двое, то она окажется меньше, чем была бы, если бы не пришёл кто-то один

- а) Верно ли первое утверждение Сергея?
- б) Верно ли утверждение Алексея?
- с) Верно ли второе утверждение Сергея?

Задача 20. Имеется 10 белых и 10 черных шаров. Вам предлагают каким-то образом разложить эти шары по двум урнам. Далее случайно выбирается одна из урн (выборы урн равновероятны), а из нее вытаскивается шар (для всех шаров, лежащих в урне, вероятности их выбора одинаковы). Если он оказывается белым, то Вы получаете приз. Как нужно разложить шары по урнам, чтобы вероятность выиграть приз была наибольшей?

Задача 21 (Диагностика 9 класс 2026). У плотника 15 серых и 15 жёлтых тротуарных плиток. Он выкладывает дорожку шириной в одну плитку и длиной 20 плиток, причём выбирает их совершенно случайно. Какова вероятность того что пятая и последняя плитки на этой дорожке окажутся одного цвета? Результат округлите до тысячных.

Тема 8

Проверка гипотез

8.1 Теоретические материалы

«Наука — это кладбище гипотез.»

Анри Пуанкаре

Одной из глобальных целей статистики является проверка гипотез. Верно ли, что средний рост мальков больше среднего роста девочек? Верно ли, что кандидата А поддерживает $k\%$ электората? Влияет ли курение на здоровье?

Гипотезой мы привыкли называть утверждение, требующее проверки. Между тем к самой формулировке гипотез нужно подходить аккуратно. Так в античность во времена холеры существовали две теории о передачи заболевания: от больного или носителя без симптомов или, так называемая, миазматическая: от гниющих органических остатков. Как следует сформулировать гипотезу, чтобы проверить какая из теорий верна?

Итак, процесс построения гипотез не прост сам по себе. Мы же обратимся к гипотезам статистическим. Дадим определение статистической гипотезе:

Определение 8.1: Статистическая гипотеза

Статистической гипотезой H (или просто гипотезой) называется любое предположение относительно вида распределения, параметров распределения или свойств закона распределения наблюдаемой в эксперименте случайной величины X .

Глобально гипотезы делятся на параметрические и непараметрические. Например гипотеза о том, что средний балл ЕГЭ по группам одинаков – параметрическая, она касается параметра распределения, в данном случае среднего.

Гипотеза же о том, что мальчики более успешны в математике чем девочки является непараметрической.

Иногда постановка вопроса влияет на то, какого рода гипотеза перед нами. Например, пусть в школе провели опрос: «Любите ли вы математику?» (да/нет). В 11 «А» из 30 учеников ответили «да» 18, а в 11 «Б» из 28 — 12. Можно ли считать, что доля любителей математики в классах одинакова? Здесь гипотеза о равенстве долей — параметрическая, но если бы мы спросили «зависит ли любовь к математике от класса?» — это уже непараметрическая гипотеза о независимости.

Определение 8.2: Параметрическая гипотеза

Любое предположение относительно параметров распределения случайной величины X называют параметрической гипотезой.

Процедура проверки гипотезы может выглядеть не очень математической. В общем виде алгоритм устроен так: предполагается некоторое утверждение, как правило заключающееся в отсутствии наблюдения значимого эффекта, которое называют нулевой или нуль гипотезой и обозначается H_0 . В противовес выбирается альтернативная гипотеза, которую обозначают H_1 или H_A , состоящая в том, что искомый эффект обнаружен.

После этого выбирается некоторое правило – статистический критерий, на основании которого мы пытаемся отвергнуть нулевую гипотезу. То есть мы никогда не пытаемся подтвердить гипотезу, напротив, в духе Базарова всегда ставим нулевую гипотезу под сомнение.

Определение 8.3: Нулевая и альтернативная гипотезы

Проверяемая гипотеза называется основной (или нулевой) и обозначается H_0 . Гипотеза, конкурирующая с H_0 , называется альтернативной и обозначается H_A .

Статистический критерий на основании выборки позволяет отвергнуть нулевую гипотезу или не позволяет, но, ввиду случайности выборки, мы можем допустить ошибки.

Ясно, что глобально мы можем совершить две разные ошибки. Например пусть тестируется новое гомеопатическое лекарственное средство. Предполагаемая нулевая гипотеза должна иметь вид

$$H_0 = \{\text{лекарство не способствует выздоровлению}\}.$$

Но если вдруг, может быть ввиду случайности, на основании тестов мы примем решение отвергнуть нулевую гипотезу в пользу альтернативы мы выпустим на рынок лекарство, пользы от которого нет. Такая ошибка называется ошибкой 1-го рода.

Определение 8.4: Ошибка 1-го рода

Ошибкой 1-го рода называется событие, состоящее в том, что гипотеза H_0 отвергается, когда она верна.

То есть ошибка первого рода заключается в нахождении эффекта там, где его нет.

Вместе с тем можно пропустить значимый эффект, не найдя статистических оснований для отвержения нулевой гипотезы. Например как в притче о мальчике, который кричал «Волки»: по выборке из ложных тревог сельчане не поверили мальчику, когда волки все же появились – не отвергли нулевую гипотезу об отсутствии эффекта.

Определение 8.5: Ошибка 2-го рода

Ошибкой 2-го рода называется событие, состоящее в том, что принимается гипотеза H_0 , когда верна гипотеза H_1 .

Таким образом мы будем считать, что возможен один из 4-х.

Какая из ошибок кажется страшнее?

Определение 8.6: Статистический критерий

Статистическим критерием проверки гипотезы H_0 называется правило, в соответствии с которым по реализации $t = \varphi(z_n)$ статистики T гипотеза H_0 принимается или отвергается.

Определение 8.7: Критическая область

Критической областью $\bar{G}$ статистического критерия называется область реализаций t статистики T , при которых гипотеза H_0 отвергается.

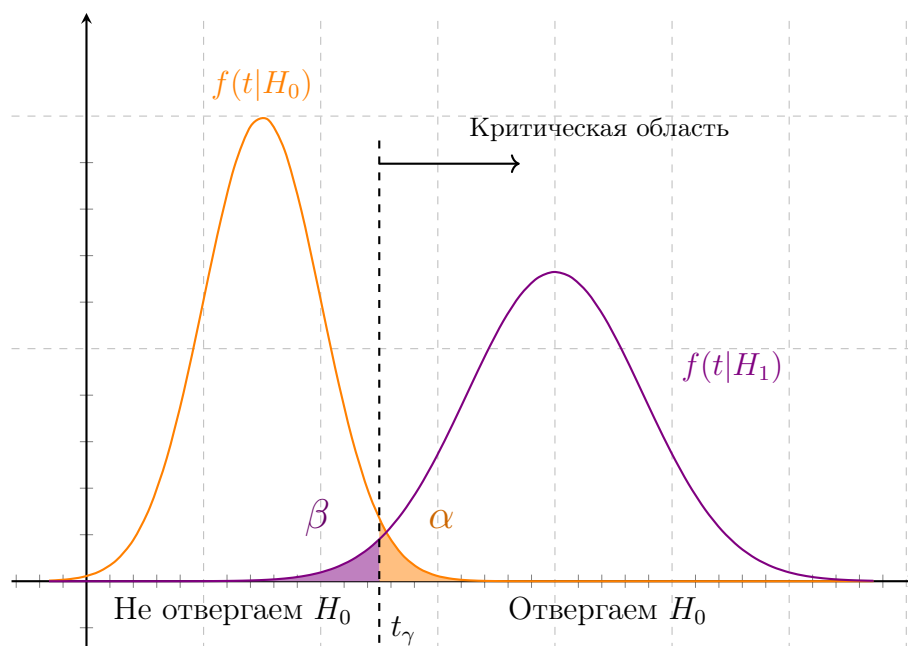
Определение 8.8: Доверительная область

Доверительной областью G статистического критерия называется область значений t статистики T , при которых гипотеза H_0 принимается.

Определение 8.9: Уровень значимости критерия

Уровнем значимости статистического критерия называется вероятность ошибки 1-го рода

$$\alpha = \mathbb{P}\{T \in G | H_0 - \text{верна}\}.$$



Ясно, что с уменьшением вероятности α ошибки 1-го рода возрастает вероятность β ошибки 2-го рода, и наоборот, т. е. при выборе критической и доверительной областей должен достигаться определенный компромисс. Поэтому вероятности ошибки 1-го рода фиксируют на некотором приемлемом значении, а критическая область выбирается таким образом, чтобы вероятность ошибки 2-го рода была минимальна.

Определение 8.10: p -value

Минимальный уровень значимости α , при котором нулевая гипотеза отвергается в пользу альтернативы обозначается как p – value.

8.2 Задачи для тренировки

Задача 1. Митя и Миша проверяли одну и ту же гипотезу одним и тем же методом по одной и той же выборке, отличались только выбранные уровни значимости – 0.02 и 0.01. Какой из них выбрал Миша, если он отверг гипотезу, а Митя не отверг?

Задача 2. У Даши есть монетка, которую она подбросила 10 раз. Орёл выпал 8 раз. Можно ли утверждать на уровне значимости $\alpha = 0.05$, что монетка честная?

Задача 3. В экзаменационном тесте по некоторому предмету предложено 10 вопросов с четырьмя вариантами ответа на каждый. Такие тесты используются для определения уровня знаний и в первую очередь нацелены на отделение неподготовленных учащихся от подготовленных. В качестве результатов теста обычно используют число правильных ответов. Вопрос в том, как определить границу числа правильных ответов, отделяющую одних (не подготовленных) учащихся от всех остальных. Другими словами где должна проходить граница между удовлетворительной и неудовлетворительной оценками?

Задача 4. Сумасшедший учёный решил занять досуг и 100 раз подбросил игральную кость. Шестёрка выпала 25 раз. Можно ли считать эту игральную кость честной?

Задача 5 (Задача о леди, дегустирующей чай.). Согласно принятой в Англии традиции чаепития, в чашку следует сначала наливать молоко, а потом - чай. Считается, что настоящая английская леди умеет отличить "правильный" чай от "неправильного". Чтобы выяснить дегустаторские качества дамы, был проведен следующий эксперимент: в течение 30 дней даме каждый день подавали пару чашек чая, в одну из которых сначала был налит чай, а в другую - молоко. Дама 21 раз верно указала "правильный" чай. Можно ли (на уровне доверия 0.95) считать её хорошим дегустатором?

Задача 6 (Эмпат). Миша и Митя решили посетить занятие по эмпатии, на котором проводился тренинг: каждый из 50 участников задавал ещё одному участнику, вопросы с бинарным ответом, например: «Ты чувствуешь себя сейчас подавленно или расслабленно?» и т.п., после чего каждый участник записывал тот ответ, который считал верным. Таким образом к концу опроса каждый записал 50 ответов, после чего интервьюируемый назвал правильные ответы. У Мити оказалось 42 верных, что вызвало шквал аплодисментов от участников, а у Миши – лишь 17, чему тот сильно расстроился. Считая уровень значимости равным 0.05, ответьте на вопросы:

- а) Была ли у участников статистически значимая причина аплодировать Мите?
- б) Была ли у Миши статистически значимая причина расстраиваться?

Задача 7. У Владика есть две на вид неразличимые монеты. Одна обыкновенная, с вероятностью $1/2$ на ней выпадает орел или решка. Вторая монетка с памятью. Если на ней выпал орел, то в следующий раз с вероятностью $2/3$ выпадет решка, а с вероятностью $1/3$ – орел. И наоборот.

Предложите эксперимент, который позволяет идентифицировать монеты с вероятностью 0.95.

Ответы к задачам

- Ответ к задаче 2.1. 15
Ответ к задаче 2.2. 0
Ответ к задаче 2.3. 0
Ответ к задаче 2.4. 7.7
Ответ к задаче 2.5. $-\frac{5}{9}$
Ответ к задаче 2.6. 145
Ответ к задаче 2.7. 45.5
Ответ к задаче 2.8. увеличится на 2
Ответ к задаче 2.9. при нахождении медианы
Ответ к задаче 2.10. 240
Ответ к задаче 2.11. 72
Ответ к задаче 2.12. не изменилось
Ответ к задаче 2.13. а, с
Ответ к задаче 2.14. b, e
Ответ к задаче 2.15. да
Ответ к задаче 2.16. Рассеянный учёный
Ответ к задаче 2.17. да
Ответ к задаче 2.18. 9, 9, 24, 24, 24
Ответ к задаче 2.19. нет, нет, 2628
Ответ к задаче 2.20. недостаточно данных (требуется рисунок)
Ответ к задаче 2.21. нет (до 5 не поднимется никогда)
Ответ к задаче 2.22. 6
Ответ к задаче 2.23. 30
Ответ к задаче 2.24. 5
Ответ к задаче 2.25. не изменилась, увеличился на 13000, увеличилась на 980
Ответ к задаче 2.26. 23
Ответ к задаче 2.27. Б
Ответ к задаче 2.28. 2, 2
Ответ к задаче 2.29. 5
Ответ к задаче 2.30. 105
Ответ к задаче 2.31. а, в
Ответ к задаче 2.32. 1, 2, 3, 100, 100
Ответ к задаче 2.33. размах не изменился, медиана увеличилась, мода увеличилась, среднее увеличилось
Ответ к задаче 2.34. 2, 3, 4, 5, 5
Ответ к задаче 2.35. а, г
Ответ к задаче 2.36. 16.5
Ответ к задаче 2.37. требуется рисунок
Ответ к задаче 2.38. да, нет
Ответ к задаче 2.39. да, нет, $38\frac{1}{7}$
Ответ к задаче 2.40. а) да; б) нет; в) $\frac{5}{6}$.
Ответ к задаче 2.41. До двух баллов.

Решения задач

