Анализ данных абитуриентов мехмата МГУ 2024

Проект для статистического практикума подготовили Кротов Владимир и Ласкин Михаил

• Цель исследования:

Провести анализ данных абитуриентов Мехмата МГУ 2024 года, чтобы выявить закономерности в распределении баллов, влияние дополнительных критериев (ГТО, индивидуальные достижения) на зачисление, а также обнаружить возможные аномалии (например, списывание на ДВИ).

• Основные задачи:

1. Анализ распределения баллов
    1. Проверить, является ли распределение баллов ЕГЭ (математика, физика, русский язык) и ДВИ нормальным.
    2. Выявить выбросы и аномалии (например, абитуриенты с низкими баллами ЕГЭ, но высокими баллами ДВИ).
2. Исследование влияния дополнительных баллов
    1. Оценить, как наличие ГТО, аттестата с отличием и других индивидуальных достижений влияет на вероятность зачисления.
    2. Определить, сколько абитуриентов прошли на бюджет благодаря дополнительным баллам.
3. Корреляционный анализ предметов
    1. Проверить наличие корреляции между баллами по математике и физике.
    2. Исследовать, существует ли зависимость между баллами по русскому языку и профильными предметами.
4. Выявление аномалий в результатах ДВИ
    1. Сравнить баллы ЕГЭ и ДВИ у абитуриентов: если у кого-то низкие баллы ЕГЭ, но высокие ДВИ, это может указывать на списывание.
    2. Проанализировать, есть ли абитуриенты с подозрительно высокими баллами ДВИ при слабых школьных результатах.

Ожидаемые результаты:

- Выявление ключевых факторов, влияющих на поступление на Мехмат МГУ.
- Обнаружение возможных несоответствий в данных (аномально высокие ДВИ при низких ЕГЭ).
- Понимание взаимосвязи между школьной подготовкой (ЕГЭ) и внутренним экзаменом (ДВИ).
- Формирование выводов о справедливости и прозрачности вступительной кампании.

Этот проект может быть полезен будущим абитуриентам, преподавателям и администрации вуза для оптимизации процесса отбора.

	№	ID конкурсного заявления	Право\пна общежитие	Сумма баллов	Баллы за\пиндивидуальные достижения	Unnamed: 5	Unnamed: 6	дви	математика	физика	русский	Статус
0	1	11101017850	Да	405	6	2	0	100	100	100	97	Зачислен
1	2	11101025540	Нет	402	6	2	0	100	100	100	94	Зачислен
2	3	11101026263	Нет	399	6	2	0	100	100	100	91	Зачислен
3	4	11101023028	Нет	398	6	2	0	100	97	96	97	Зачислен
4	5	11101018150	Нет	396	6	2	2	100	100	100	86	Зачислен

Для простоты работы со столбцами переименуем их так, чтобы с ними было легче работать. Не будем отказываться полностью от русского языка, но сократим название каждого столбца до одного слова. Всегда удобнее работать с типом данных bool, чем с типом данных string. Изменим тип данных в столбцах типа "Да/Нет" на логический:

	№	общежитие	сумма	аттестат	сочинение	ГТО	ДВИ	математика	физика	русский	Статус
0	1	1	405	6	2	0	100	100	100	97	1
1	2	0	402	6	2	0	100	100	100	94	1
2	3	0	399	6	2	0	100	100	100	91	1
3	4	0	398	6	2	0	100	97	96	97	1
4	5	0	396	6	2	2	100	100	100	86	1

	№	общежитие	сумма	аттестат	сочинение	ГТО	ДВИ	математика	физика	русский	Статус
count	1292.000000	1292.000000	1292.000000	1292.000000	1292.000000	1292.000000	1292.000000	1292.000000	1292.000000	1292.000000	1292.000000
mean	646.500000	0.470588	341.261610	2.851393	1.975232	0.656347	71.977554	91.131579	87.484520	85.184985	0.264706
std	373.112584	0.499327	28.406144	2.997477	0.221269	0.939460	16.318204	6.833056	9.628767	8.557311	0.441347
min	1.000000	0.000000	245.000000	0.000000	0.000000	0.000000	50.000000	64.000000	55.000000	60.000000	0.000000
25%	323.750000	0.000000	324.000000	0.000000	2.000000	0.000000	60.000000	86.000000	80.000000	81.000000	0.000000
50%	646.500000	0.000000	340.000000	0.000000	2.000000	0.000000	70.000000	92.000000	90.000000	86.000000	0.000000
75%	969.250000	1.000000	361.000000	6.000000	2.000000	2.000000	85.000000	97.000000	96.000000	91.000000	1.000000
max	1292.000000	1.000000	408.000000	6.000000	2.000000	2.000000	100.000000	100.000000	100.000000	100.000000	1.000000

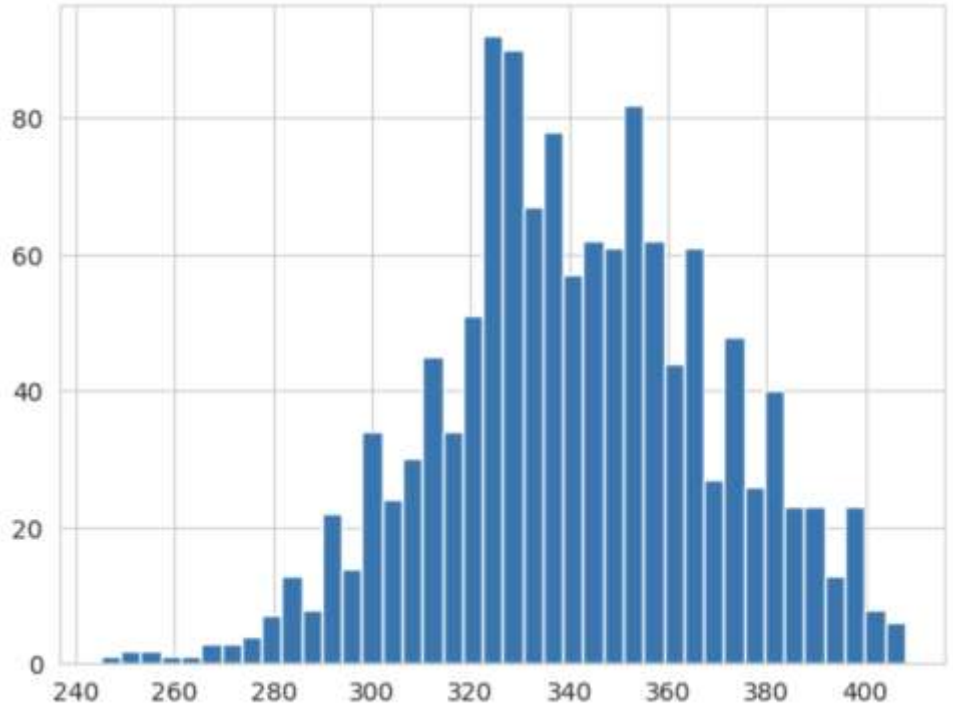
```
entrants.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1292 entries, 0 to 1291
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   №            1292 non-null   int64
 1   общежитие    1292 non-null   int64
 2   сумма        1292 non-null   int64
 3   аттестат     1292 non-null   int64
 4   сочинение    1292 non-null   int64
 5   ГТО          1292 non-null   int64
 6   ДВИ          1292 non-null   int64
 7   математика   1292 non-null   int64
 8   физика       1292 non-null   int64
 9   русский      1292 non-null   int64
 10  Статус       1292 non-null   int64
dtypes: int64(11)
memory usage: 111.2 KB
```

Оценим общее описание всех параметров данных
Средний балл:
Математика – 91
Физика – 87
Русский – 85
ДВИ – 72

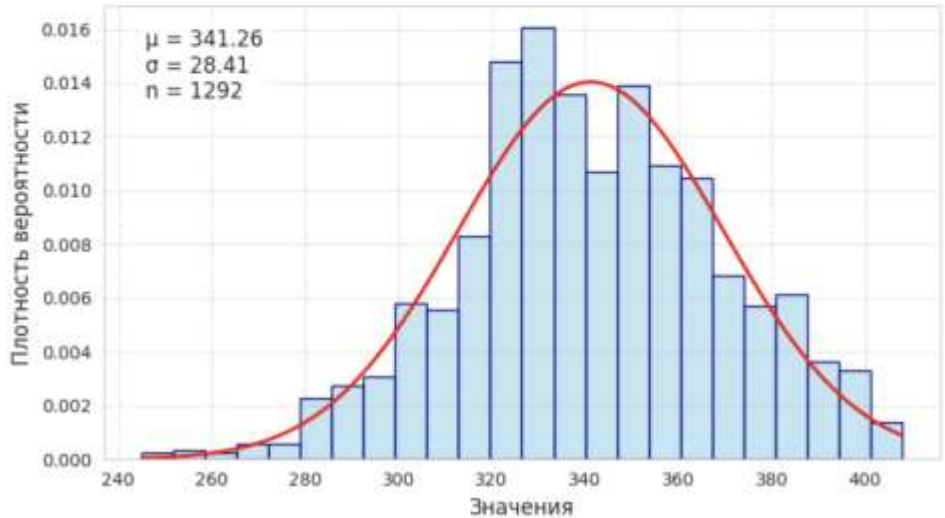
Всего 1292 абитуриента

- Первое, что стоит сделать — посмотреть, как выглядит распределение суммы баллов на гистограмме. Мы увидим, какую структуру имею т данные.

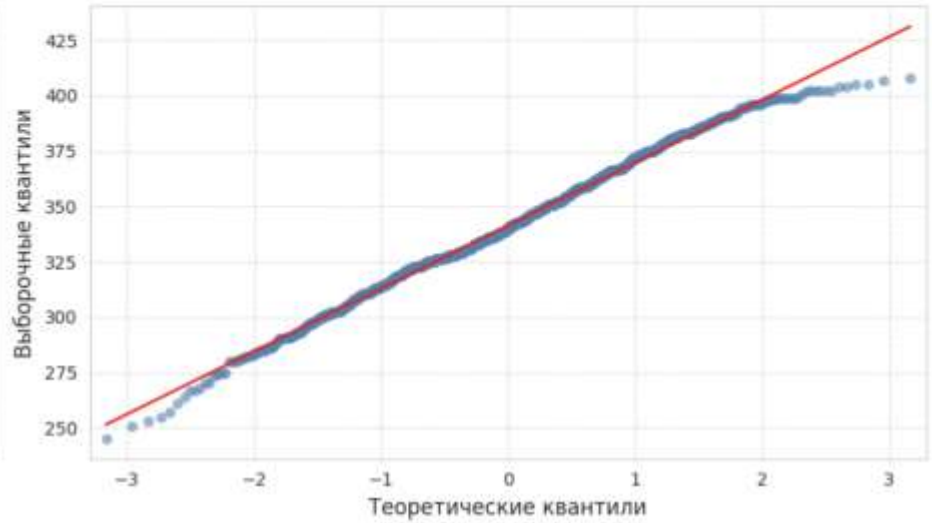


Проверка нормальности распределения: сумма

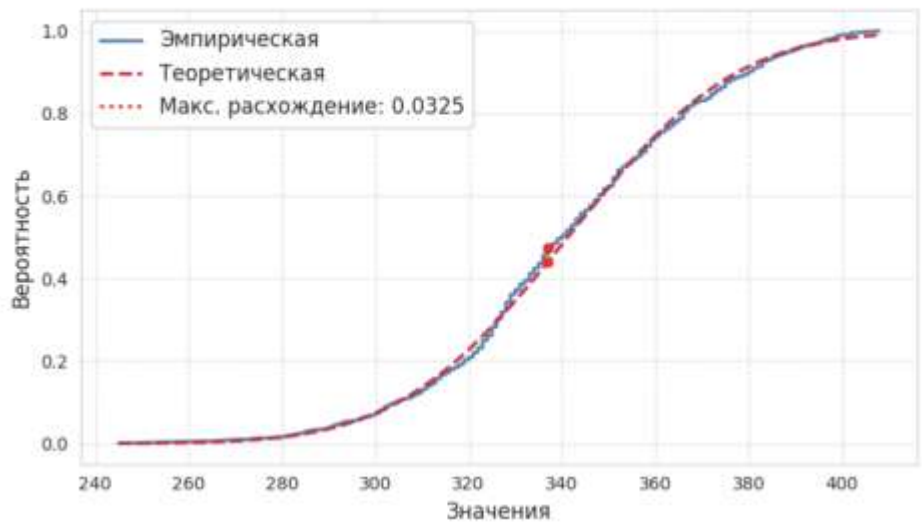
Гистограмма распределения



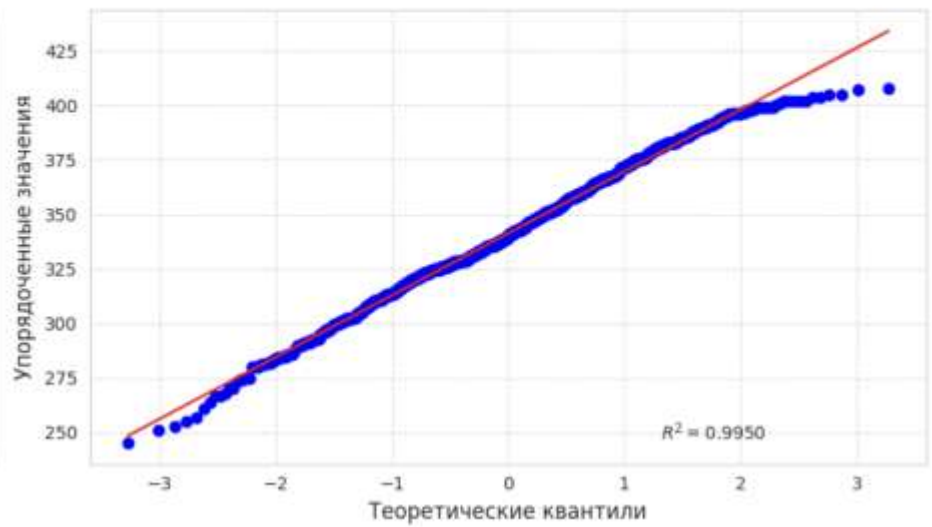
Квантиль-квантиль график (Q-Q plot)



Сравнение кумулятивных функций распределения

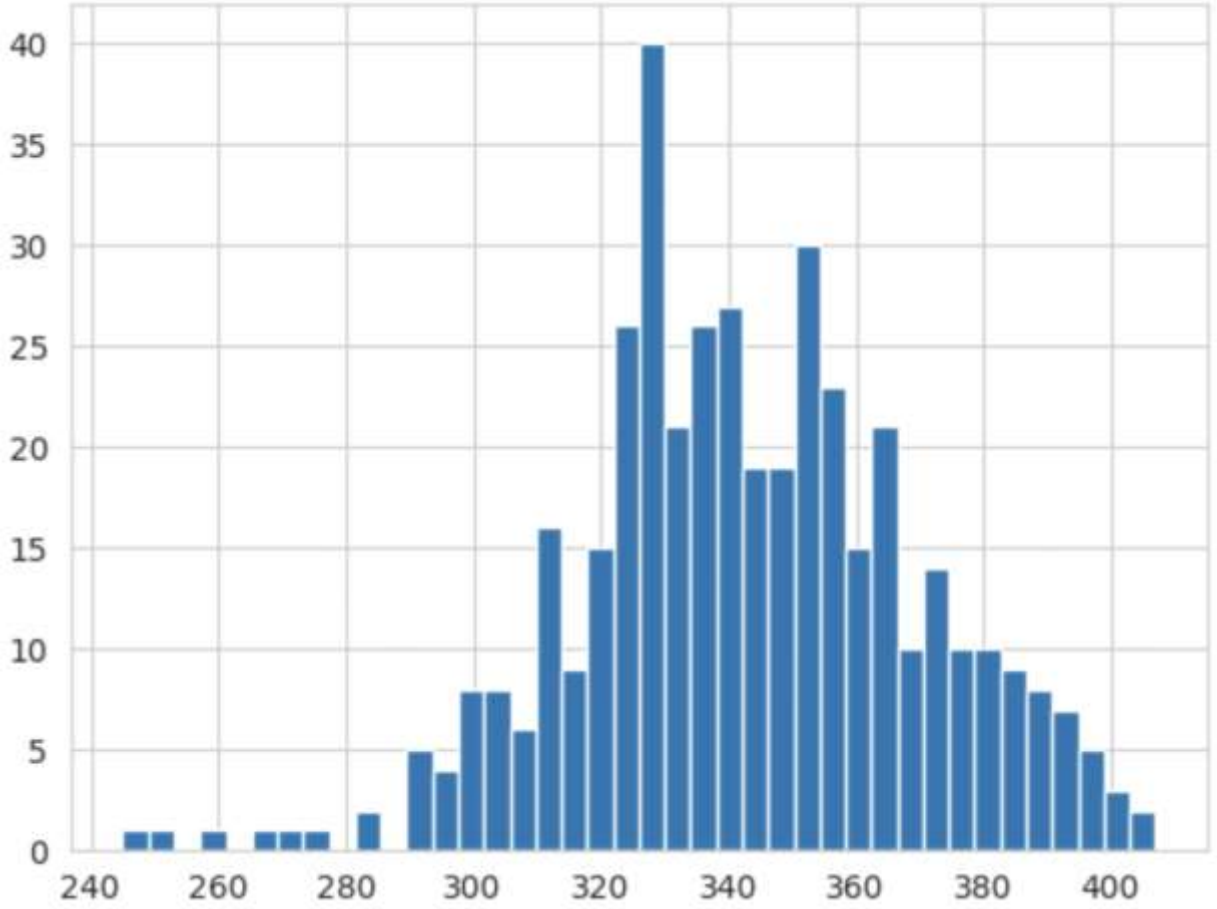


Вероятностный график (ProbPlot)



- === Результаты теста Колмогорова-Смирнова ===
- Статистика K-S: 0.0325
- P-значение: 0.1269
- Размер выборки: 1292
- Оценка асимметрии: -0.0984
- Оценка эксцесса: -0.1641
- Вывод: распределение не отличается от нормального (p > 0.05)
- Проверим, такое же распределение получается у студентов, сдавших успешно ГТО? Может быть, ГТО существенно повлияло на их поступление, хотя они менее или более талантливы, чем другие студенты?

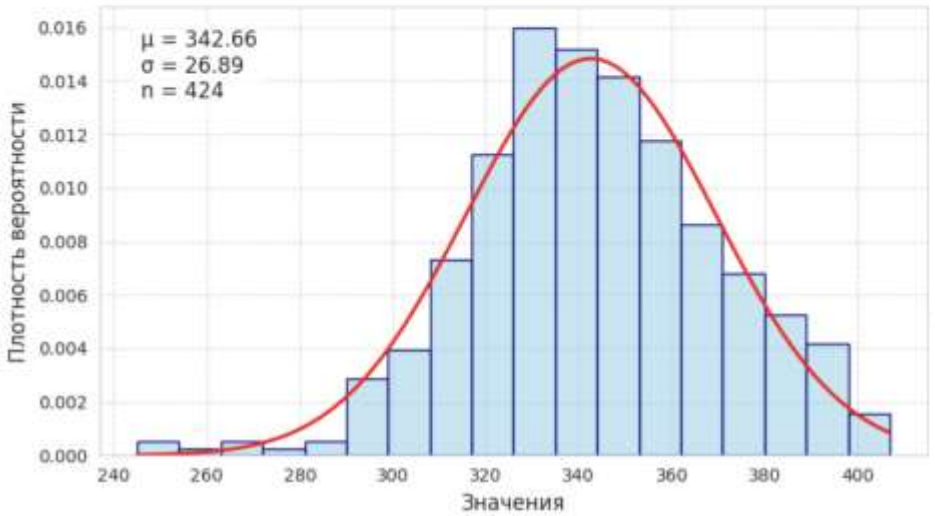
Распределение баллов у студентов, успешно сдавших ГТО



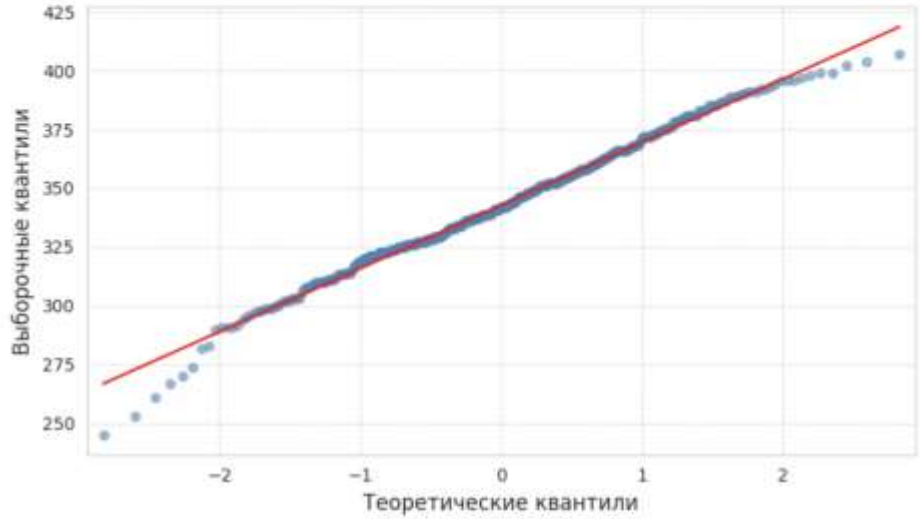
Видим, что баллы студентов, которые сдали ГТО, имеют нормальное распределение. Хочется сделать вывод, что их результаты более предсказуемы и имеют стабильное распределение. В целом, ГТО сдают более физически подготовленные студенты, что действительно может коррелировать с дисциплинированностью и стабильностью в учебе.

Проверка нормальности распределения: сумма

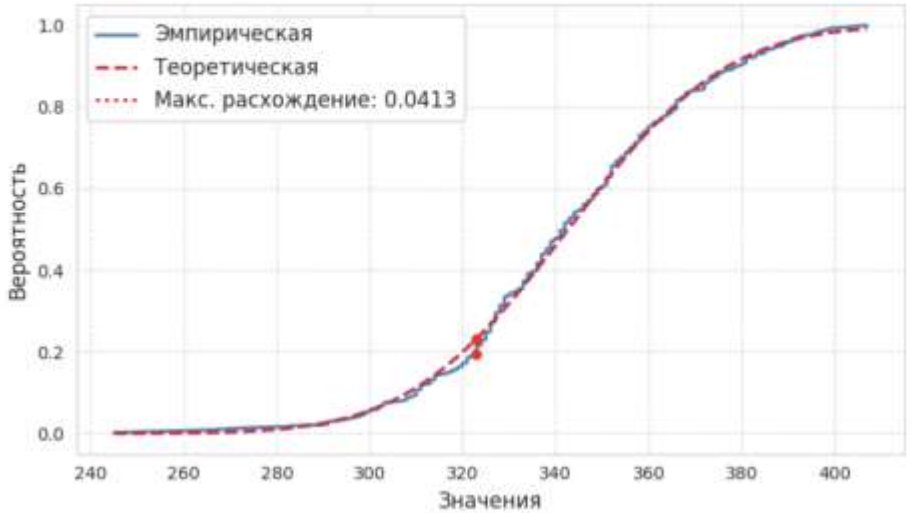
Гистограмма распределения



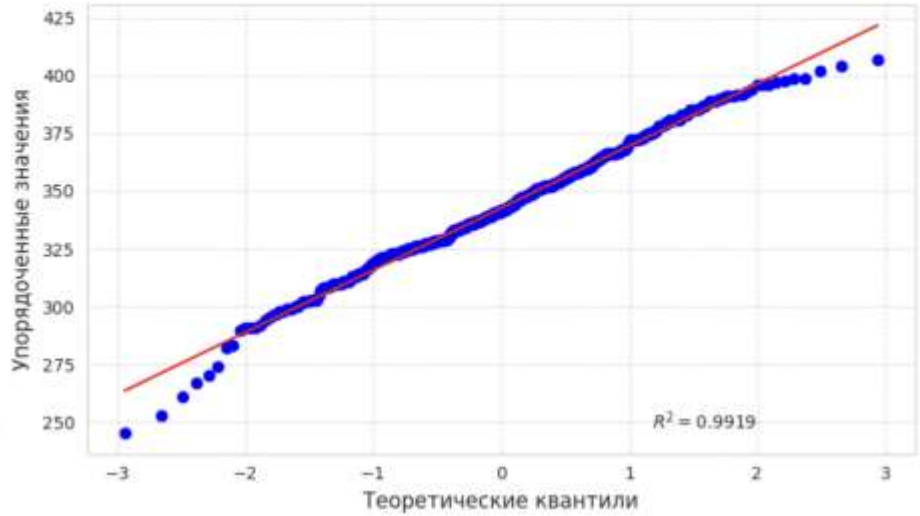
Квантиль-квантиль график (Q-Q plot)



Сравнение кумулятивных функций распределения



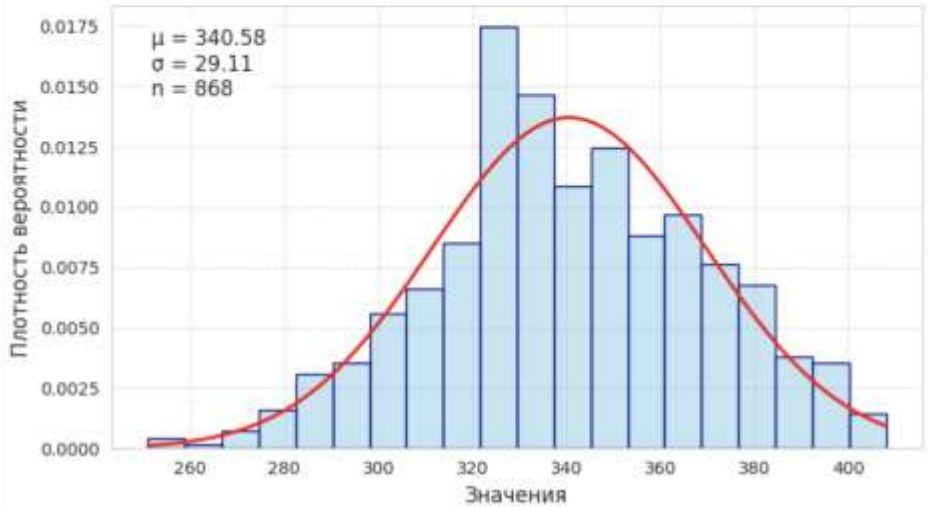
Вероятностный график (ProbPlot)



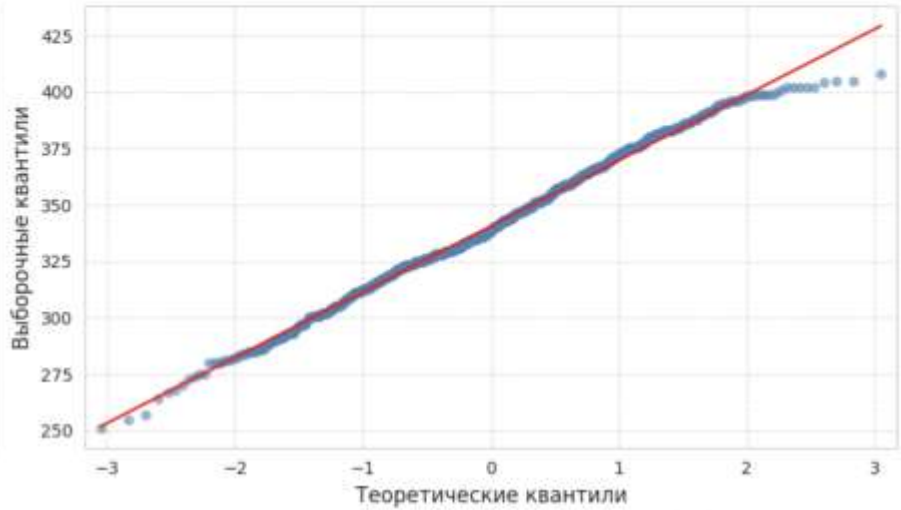
- === Результаты теста Колмогорова-Смирнова ===
- Статистика K-S: 0.0413
- P-значение: 0.4519
- Размер выборки: 424
- Оценка асимметрии: -0.1842
- Оценка эксцесса: 0.2641
- Вывод: распределение не отличается от нормального (p > 0.05)
- Как мы видим, распределение очень похоже, имеет тот же максимум. Значит, наша гипотеза неверна. А что насчет несдавших ГТО?

Проверка нормальности распределения: сумма

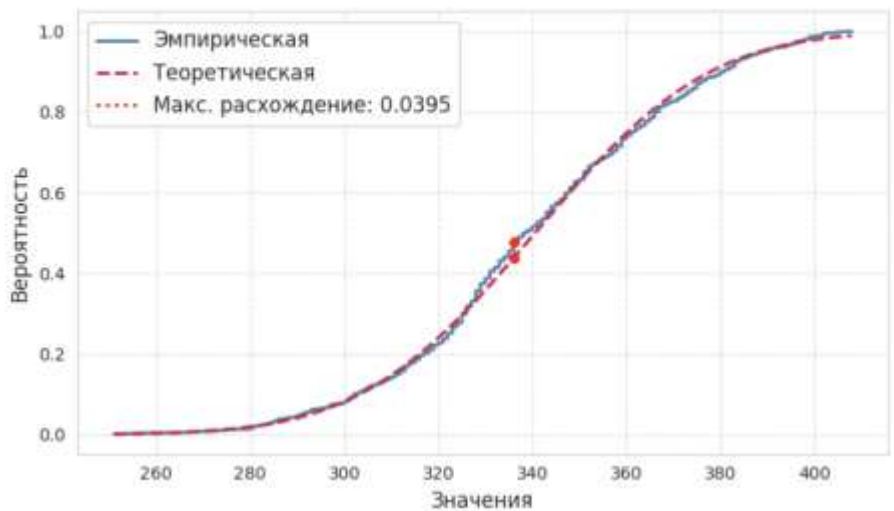
Гистограмма распределения



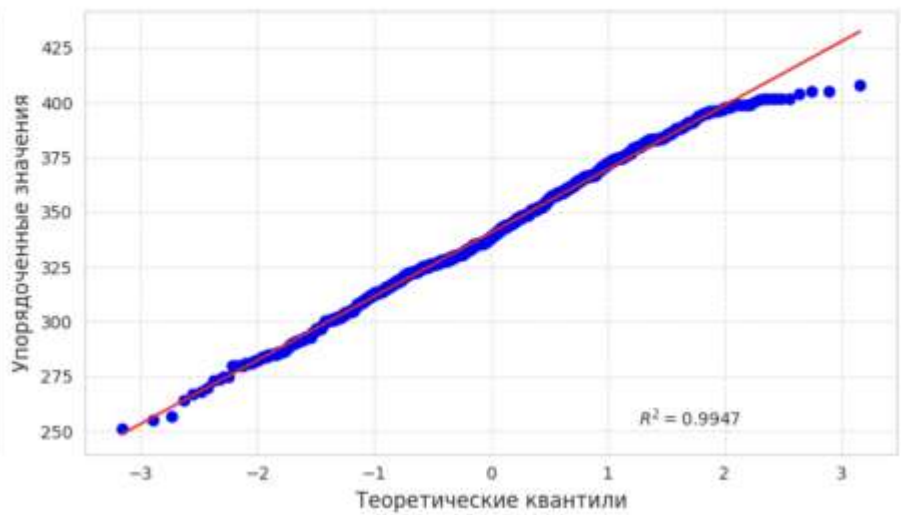
Квантиль-квантиль график (Q-Q plot)



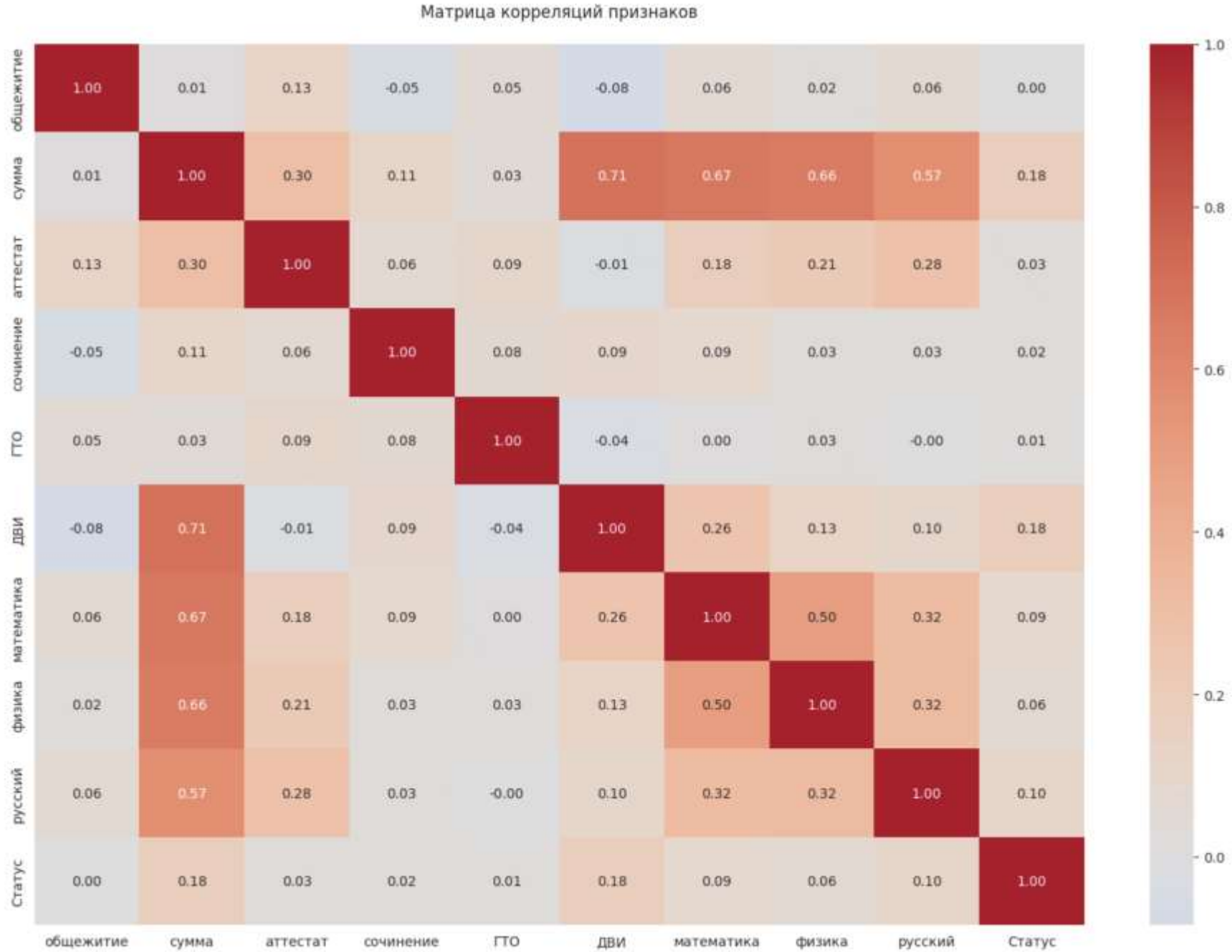
Сравнение кумулятивных функций распределения



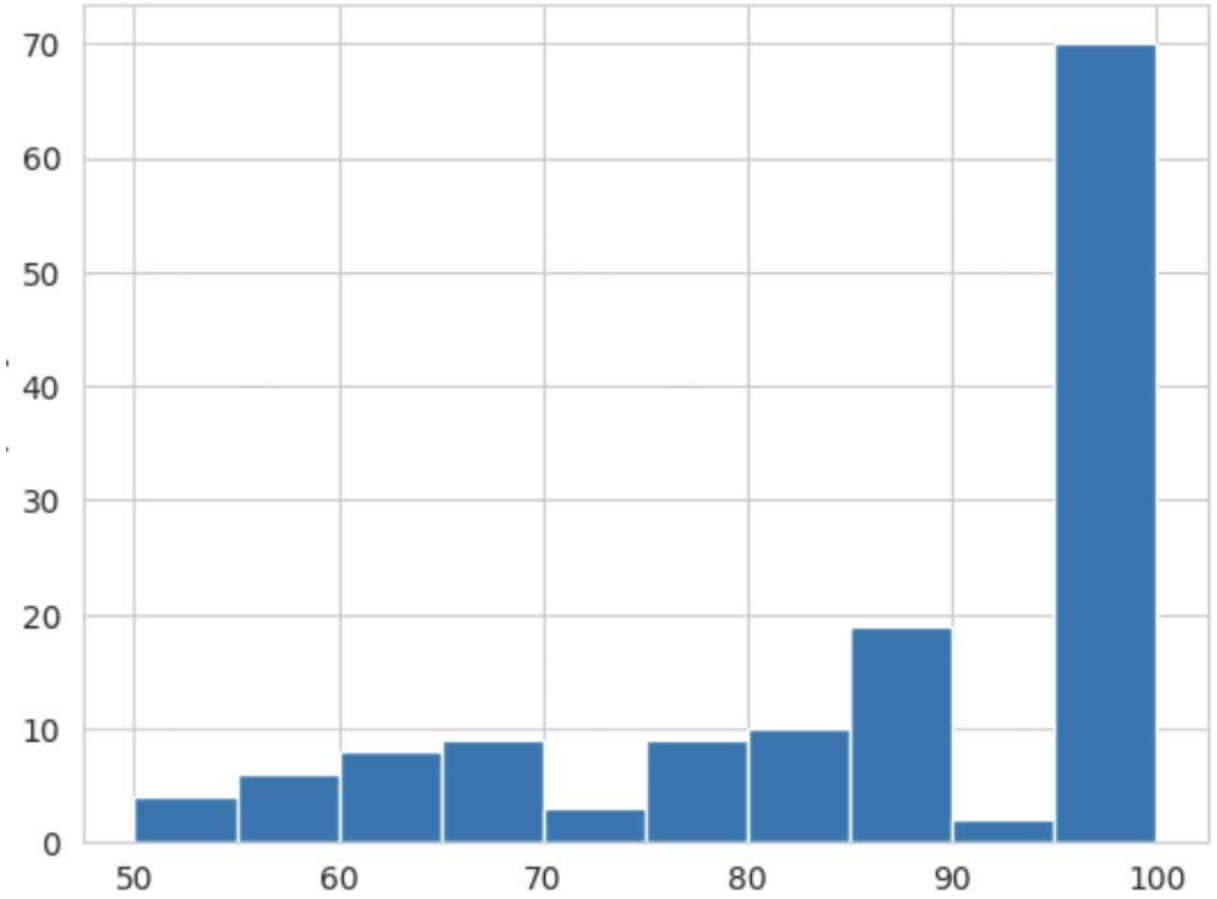
Вероятностный график (ProbPlot)



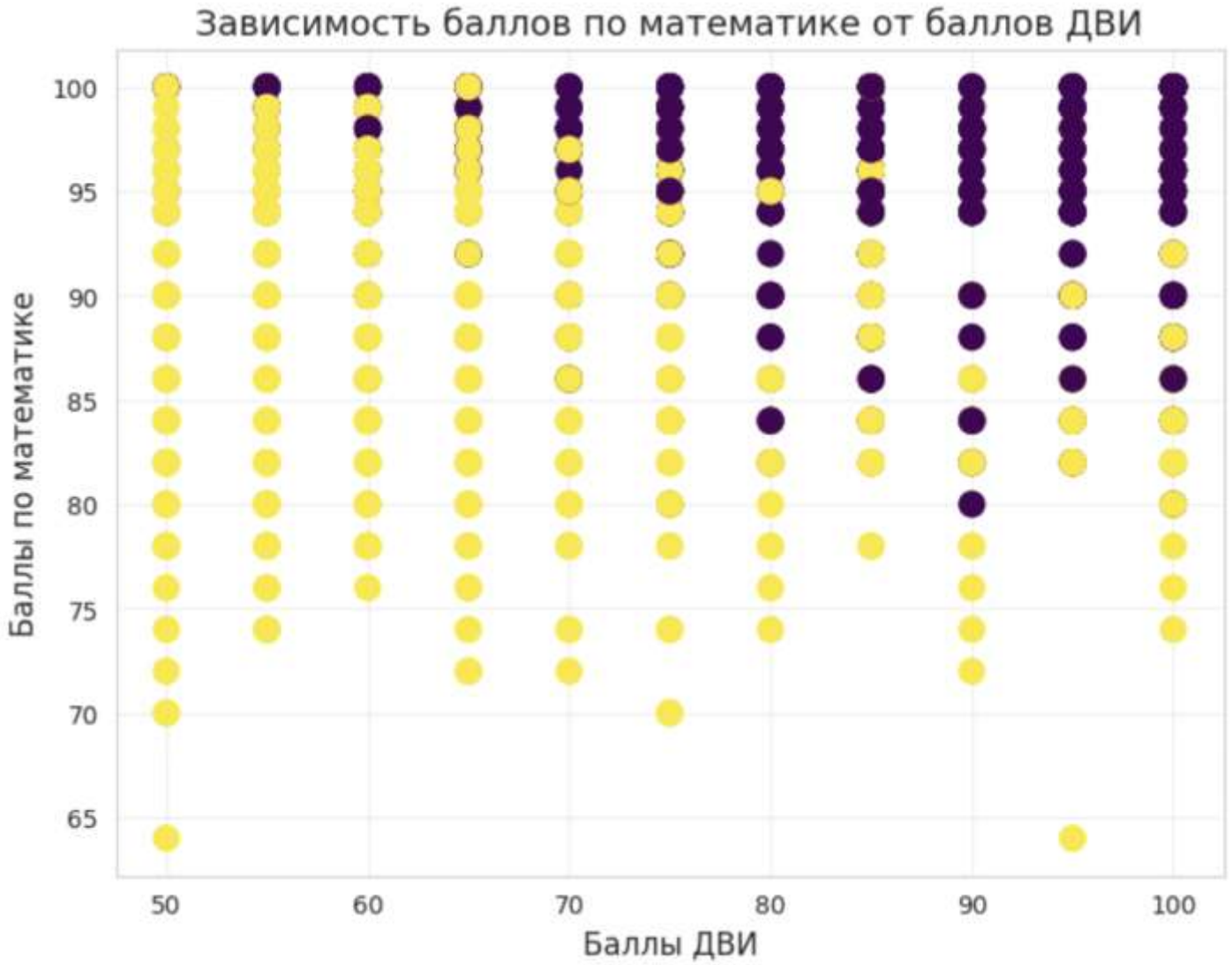
- === Результаты теста Колмогорова-Смирнова ===
- Статистика K-S: 0.0395
- P-значение: 0.1302
- Размер выборки: 868
- Оценка асимметрии: -0.0551
- Оценка эксцесса: -0.3279
- Вывод: распределение не отличается от нормального (p > 0.05)
- Видим, что у студентов, которые не сдавали ГТО, баллы тоже имеют нормальное распределение, а различия параметров незначительны. Значит, K-S тест отреагировал на тонкости распределения (скошенность, пики, "тяжелые" хвосты), а сдача ГТО на самом деле не связана с баллами за экзамен.


Матрица корреляций признаков
общежитие
сумма
аттестат
сочинение
ГТО
ДВИ
математика
физика
русский
Статус
1.00 0.01 0.13 -0.05 0.05 -0.08 0.06 0.02 0.06 0.00
0.01 1.00 0.30 0.11 0.03 0.71 0.67 0.66 0.57 0.18
0.13 0.30 1.00 0.06 0.09 -0.01 0.18 0.21 0.28 0.03
-0.05 0.11 0.06 1.00 0.08 0.09 0.09 0.03 0.03 0.02
0.05 0.03 0.09 0.08 1.00 -0.04 0.00 0.03 -0.00 0.01
-0.08 0.71 -0.01 0.09 -0.04 1.00 0.26 0.13 0.10 0.18
0.06 0.67 0.18 0.09 0.00 0.26 1.00 0.50 0.32 0.09
0.02 0.66 0.21 0.03 0.03 0.13 0.50 1.00 0.32 0.06
0.06 0.57 0.28 0.03 -0.00 0.10 0.32 0.32 1.00 0.10
0.00 0.18 0.03 0.02 0.01 0.18 0.09 0.06 0.10 1.00
общежитие
сумма
аттестат
сочинение
ГТО
ДВИ
математика
физика
русский
Статус
1.0
0.8
0.6
0.4
0.2
0.0

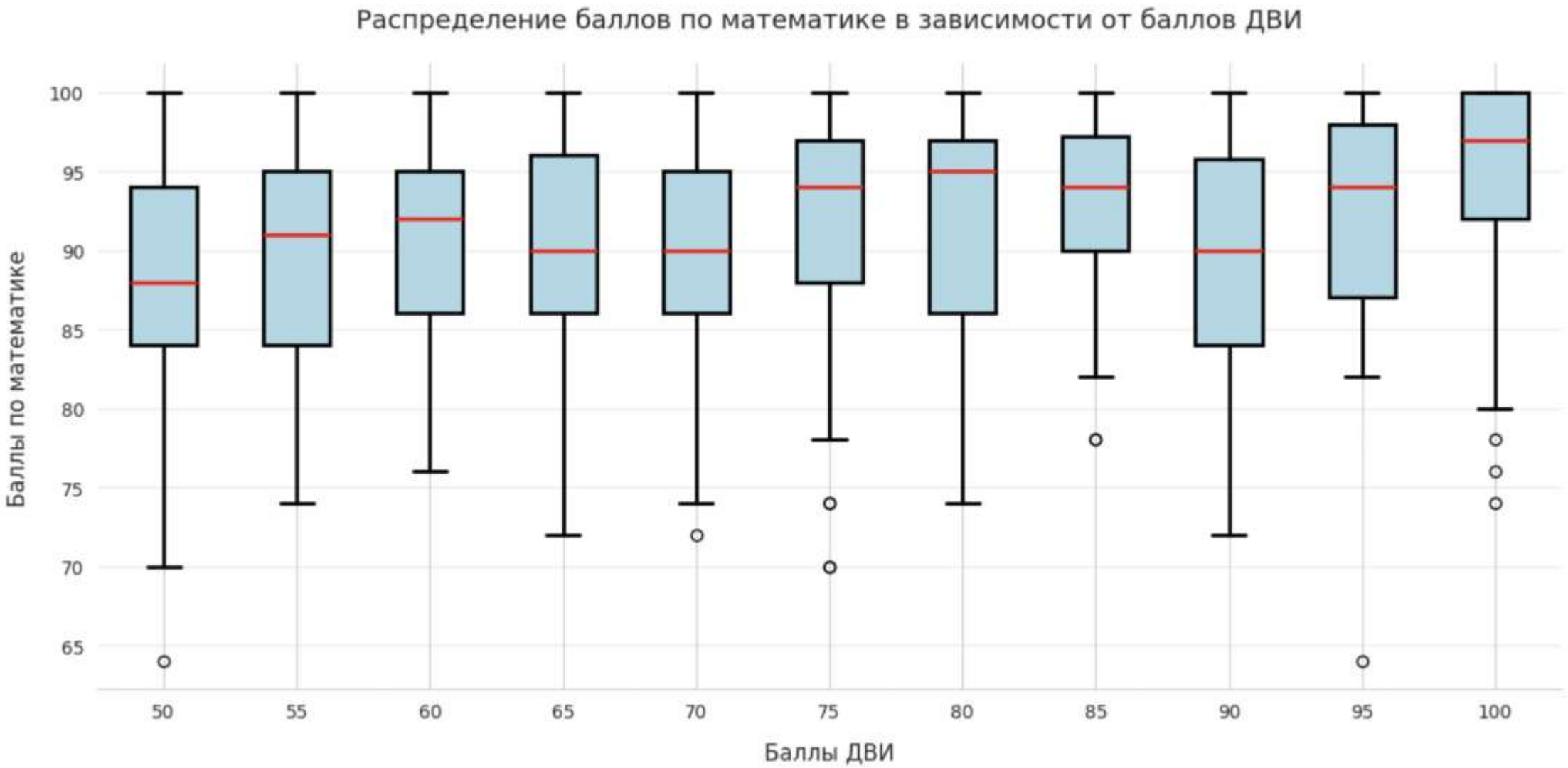
Посмотрим на сто балльников по ДВИ математике



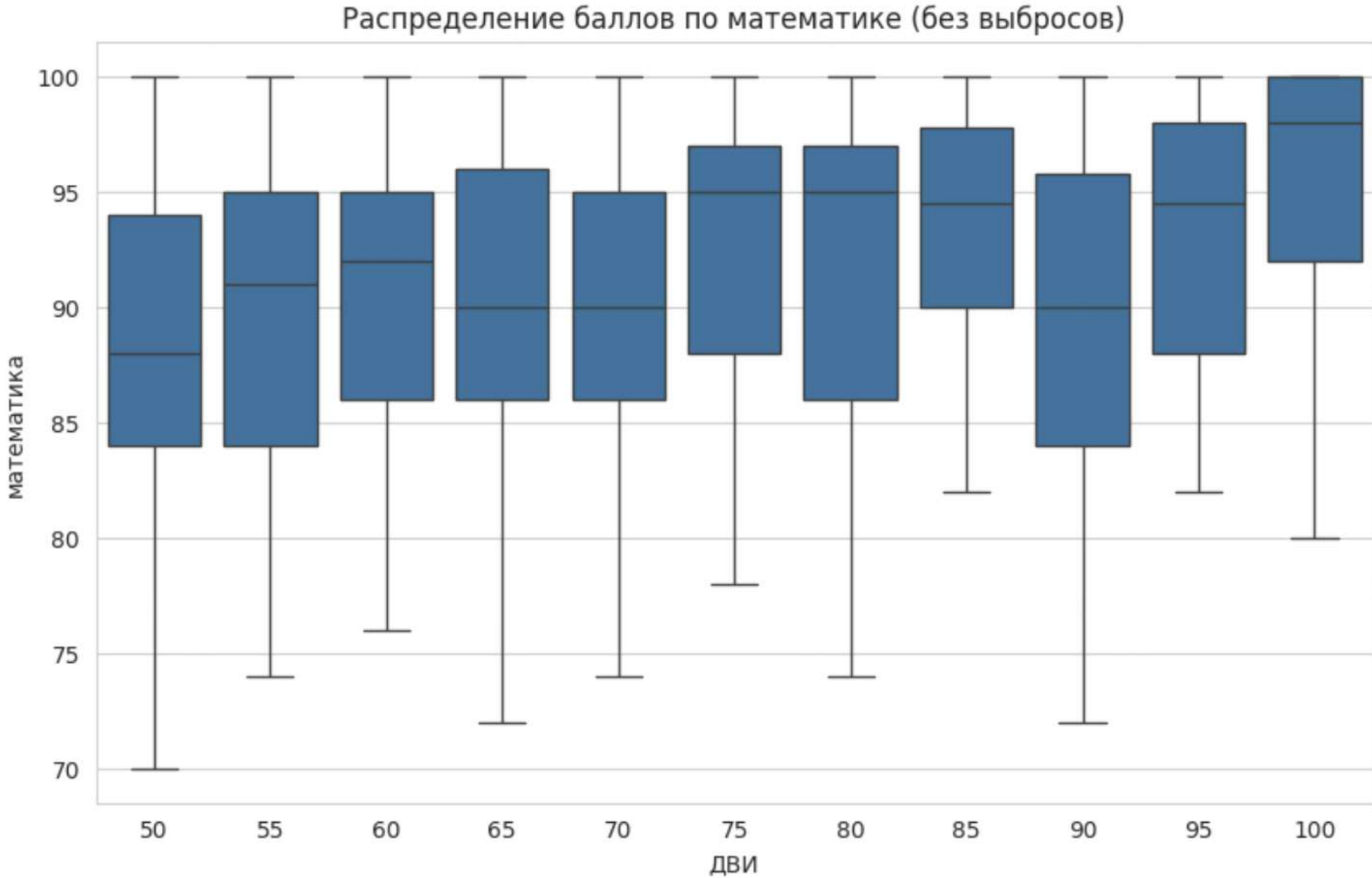
- Выдвинем следующую гипотезу:
- студенты, успешно сдавшие ЕГЭ по математике, должны были хорошо сдать и ДВИ. Если это так и если мы отметим всех студентов в координатах "Балл за ДВИ" - "Балл за ЕГЭ по математике", мы должны увидеть скопление в области высоких баллов по обеим осям. Изобразим такой поточечный график:

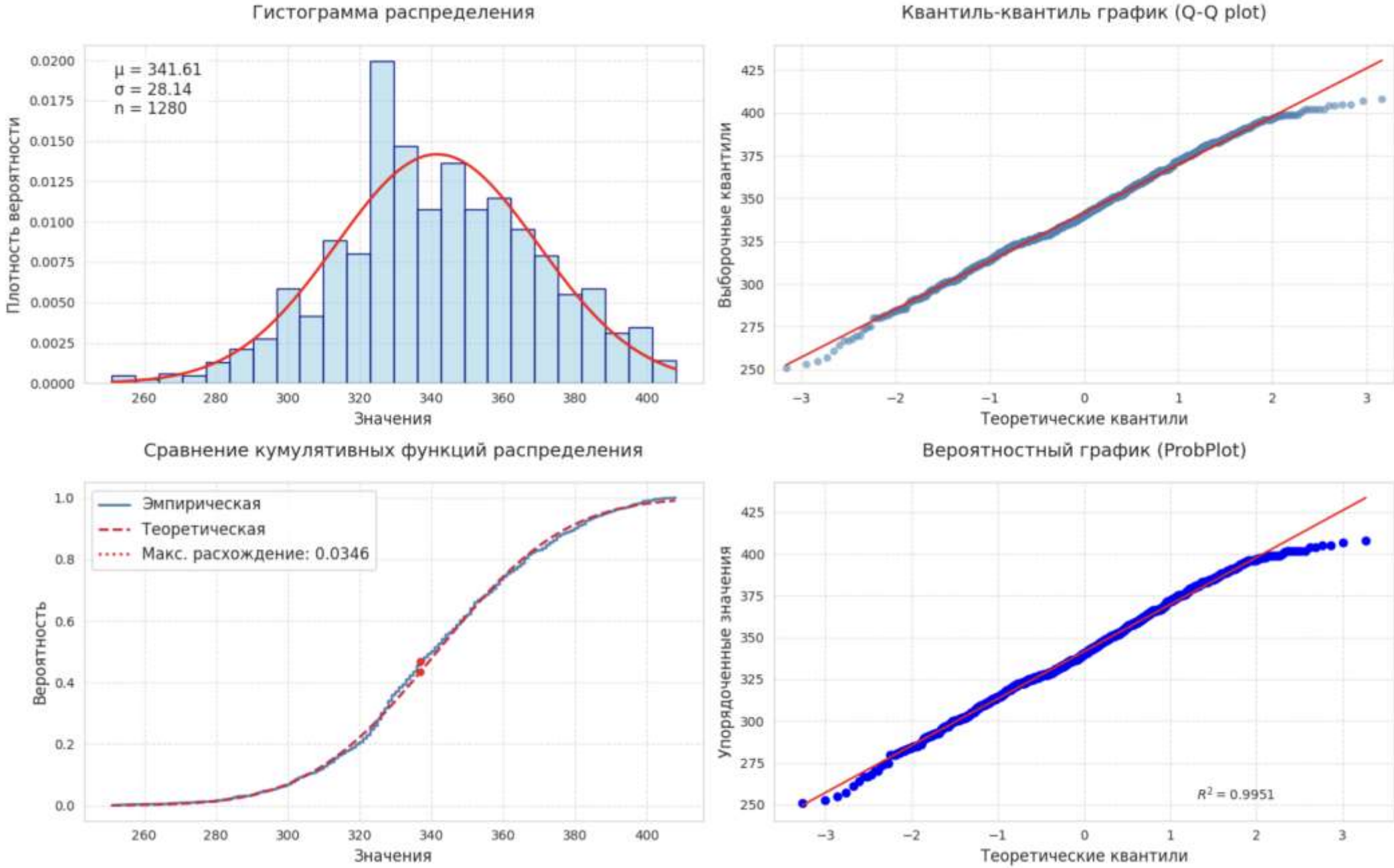


- Кажется, что распределение баллов по математике при каждом фиксированном балле за ДВИ почти равномерное в области "Балл за математику > 75". Но так ли это в действительности? Может быть, все дело в особенности отображения поточечного графика?
- Нарисуем другой тип графика, так называемые "свечи":



Удалим тех, кто вероятнее всего списывал ДВИ

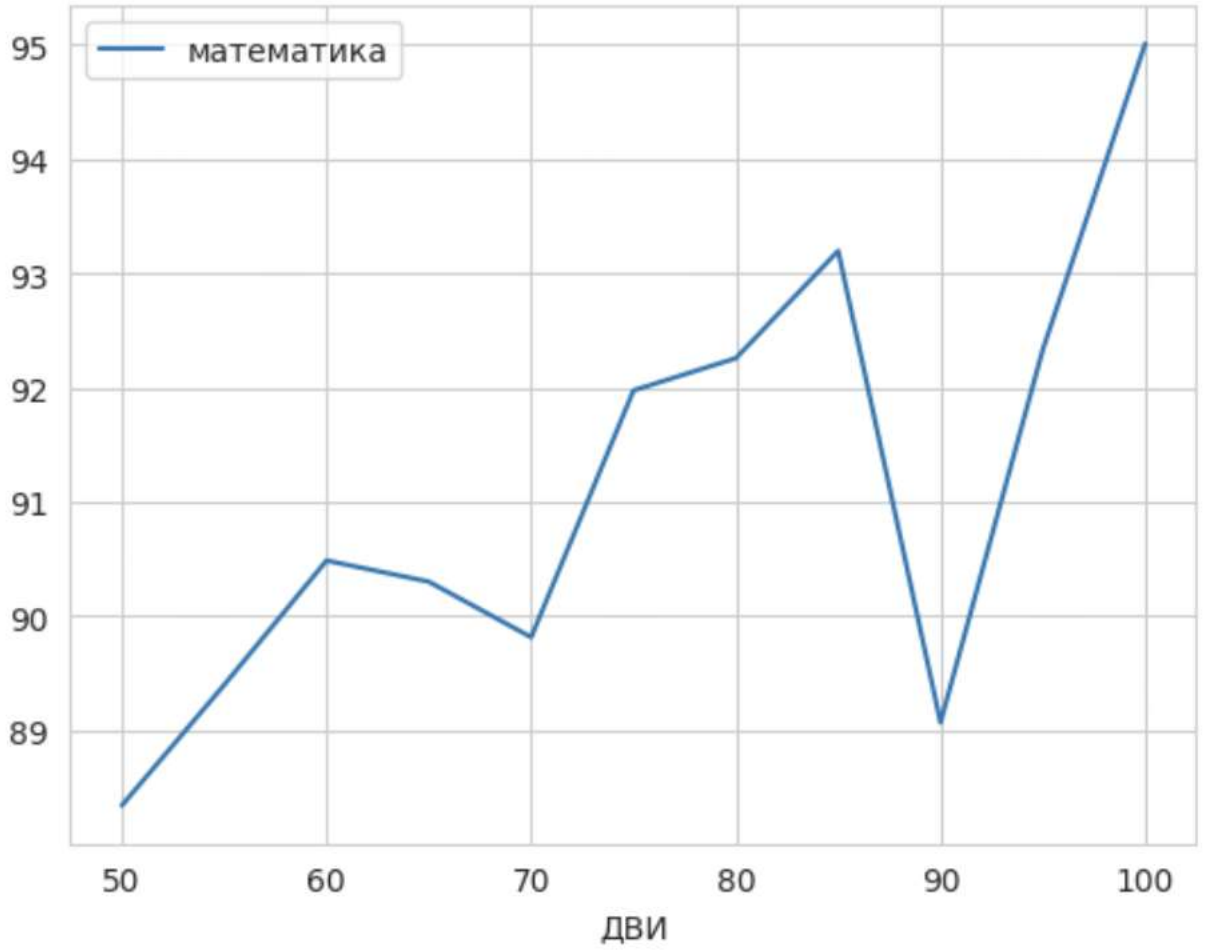

Распределение баллов по математике (без выбросов)
математика
100
95
90
85
80
75
70
50
55
60
65
70
75
80
85
90
95
100
ДВИ


Гистограмма распределения
μ = 341.61
σ = 28.14
n = 1280
Плотность вероятности
Значения
Квантиль-квантиль график (Q-Q plot)
Выборочные квантили
Теоретические квантили
Сравнение кумулятивных функций распределения
Эмпирическая
Теоретическая
Макс. расхождение: 0.0346
Вероятность
Значения
Вероятностный график (ProbPlot)
Упорядоченные значения
Теоретические квантили
$R^2 = 0.9951$

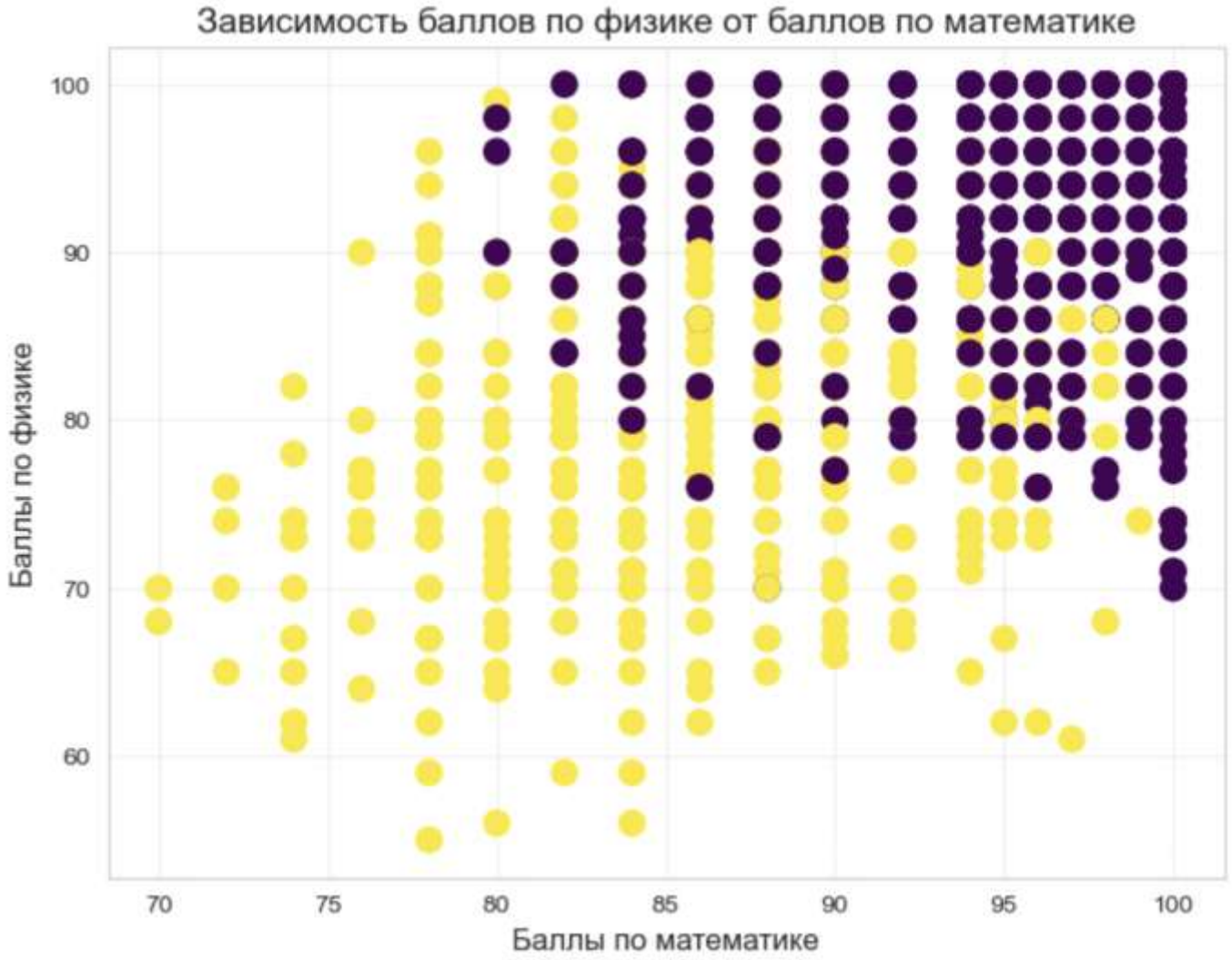
- === Результаты теста Колмогорова-Смирнова ===
- Статистика K-S: 0.0346
- P-значение: 0.0909
- Размер выборки: 1280
- Оценка асимметрии: -0.0696
- Оценка эксцесса: -0.2124
- Вывод: распределение не отличается от нормального ($p > 0.05$)

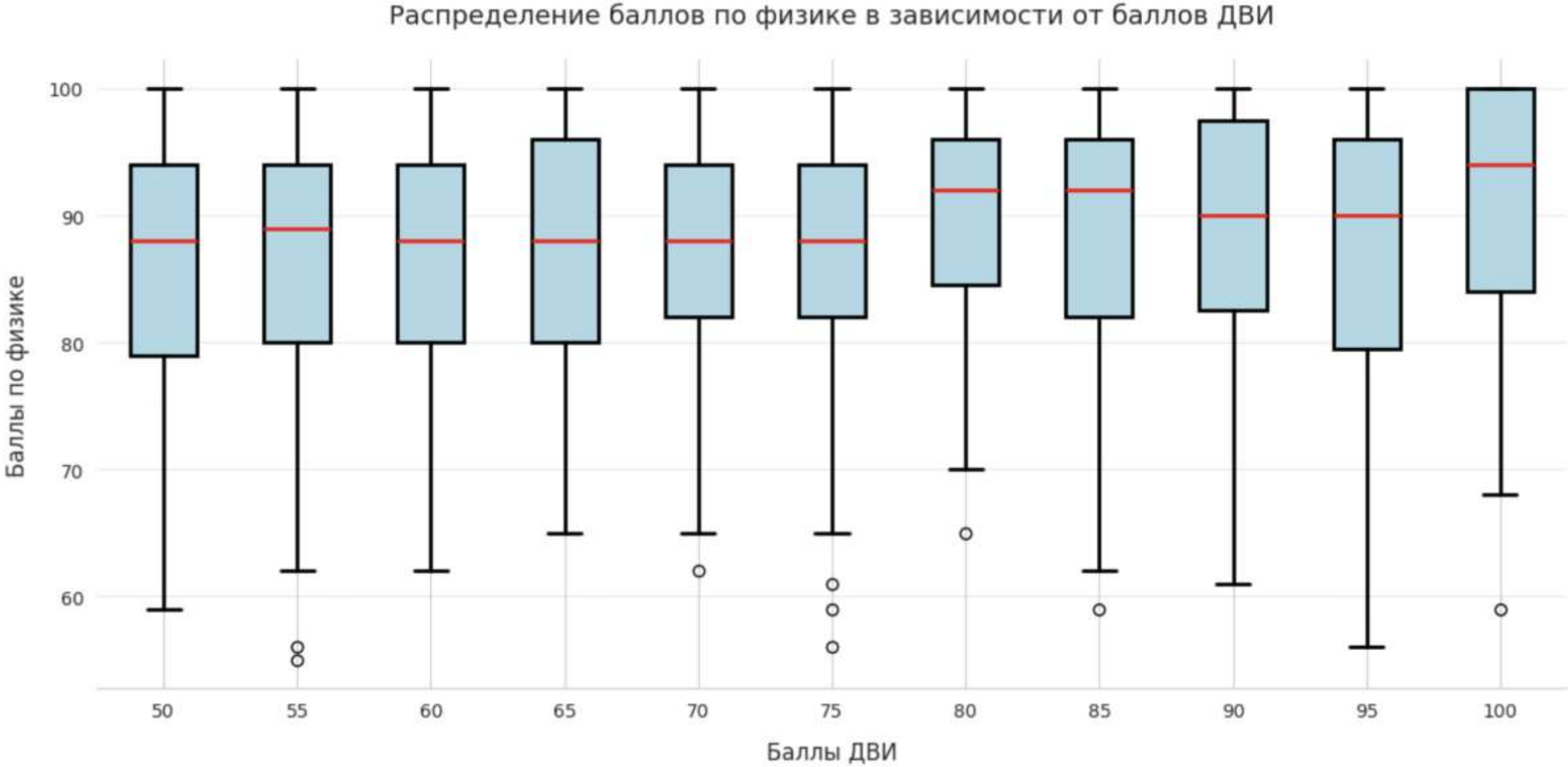
- Мы действительно наблюдаем смещение медианного значения балла ЕГЭ по математике ближе к 100 при увеличении баллов за ДВИ. Это подтверждает нашу гипотезу. Проверим это, применив усреднение:

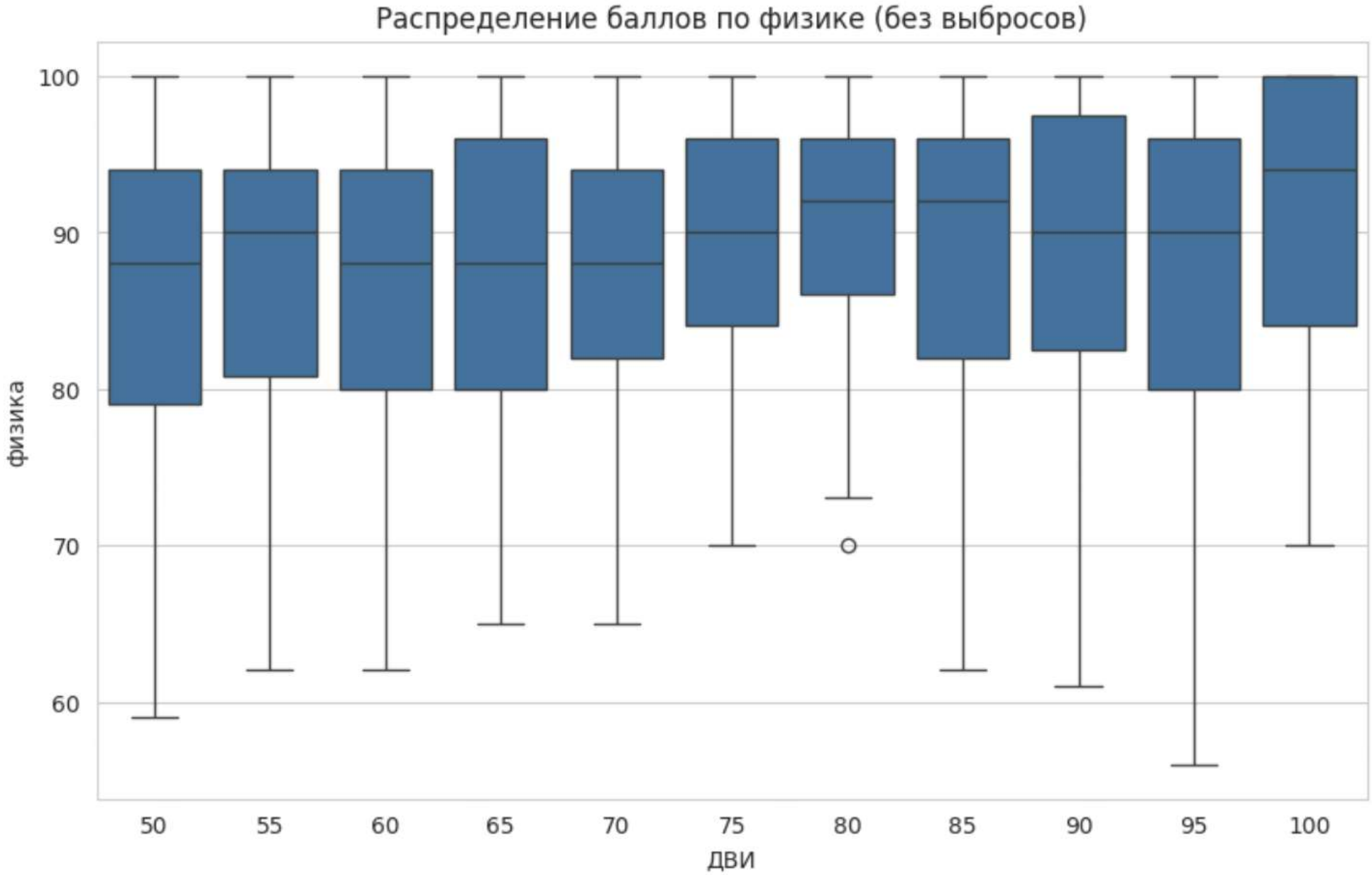
ДВИ	математика
50	88.345912
55	89.400000
60	90.487437
65	90.302521
70	89.817308
75	91.975207
80	92.256757
85	93.195312
90	89.071429
95	92.333333
100	95.006536

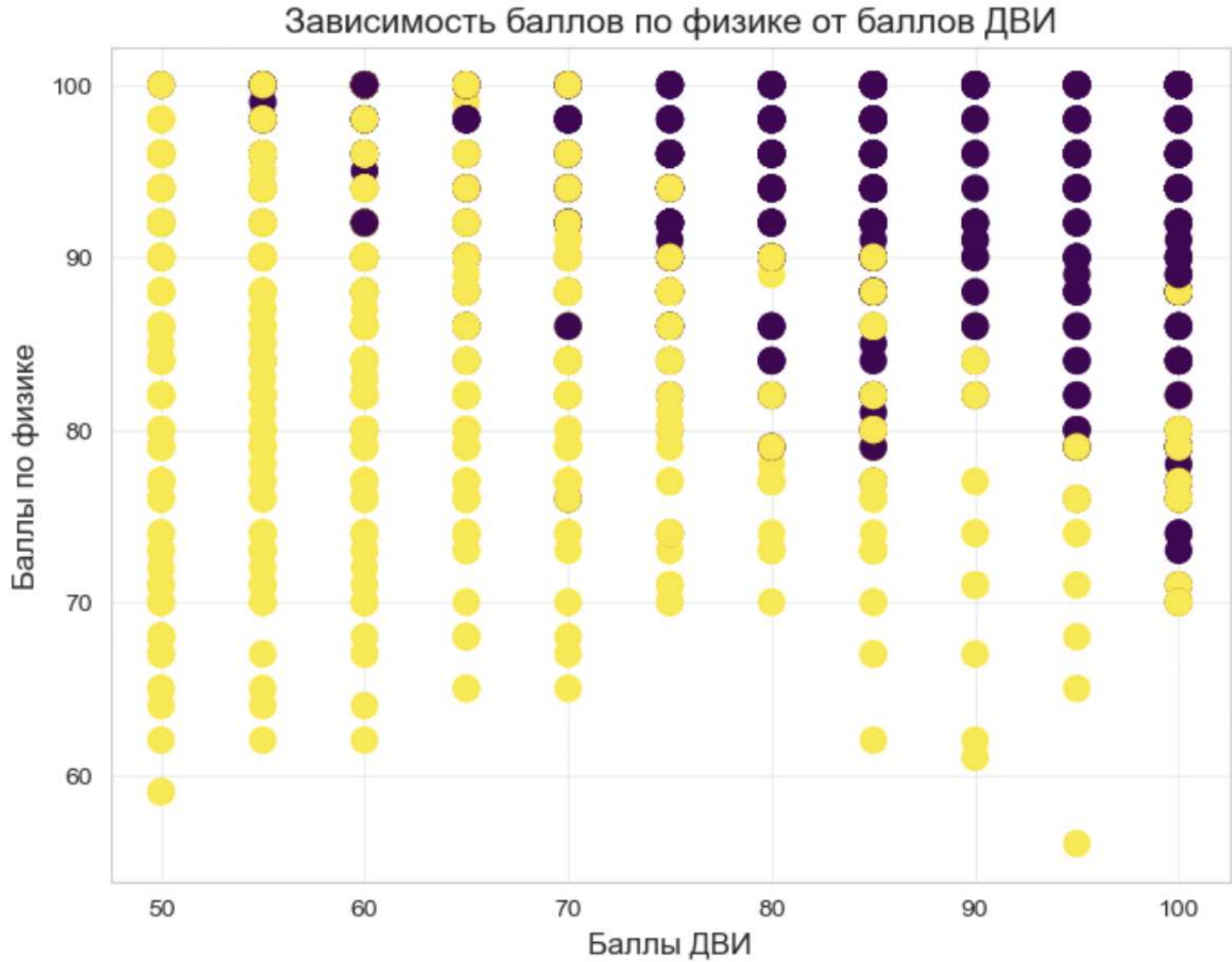


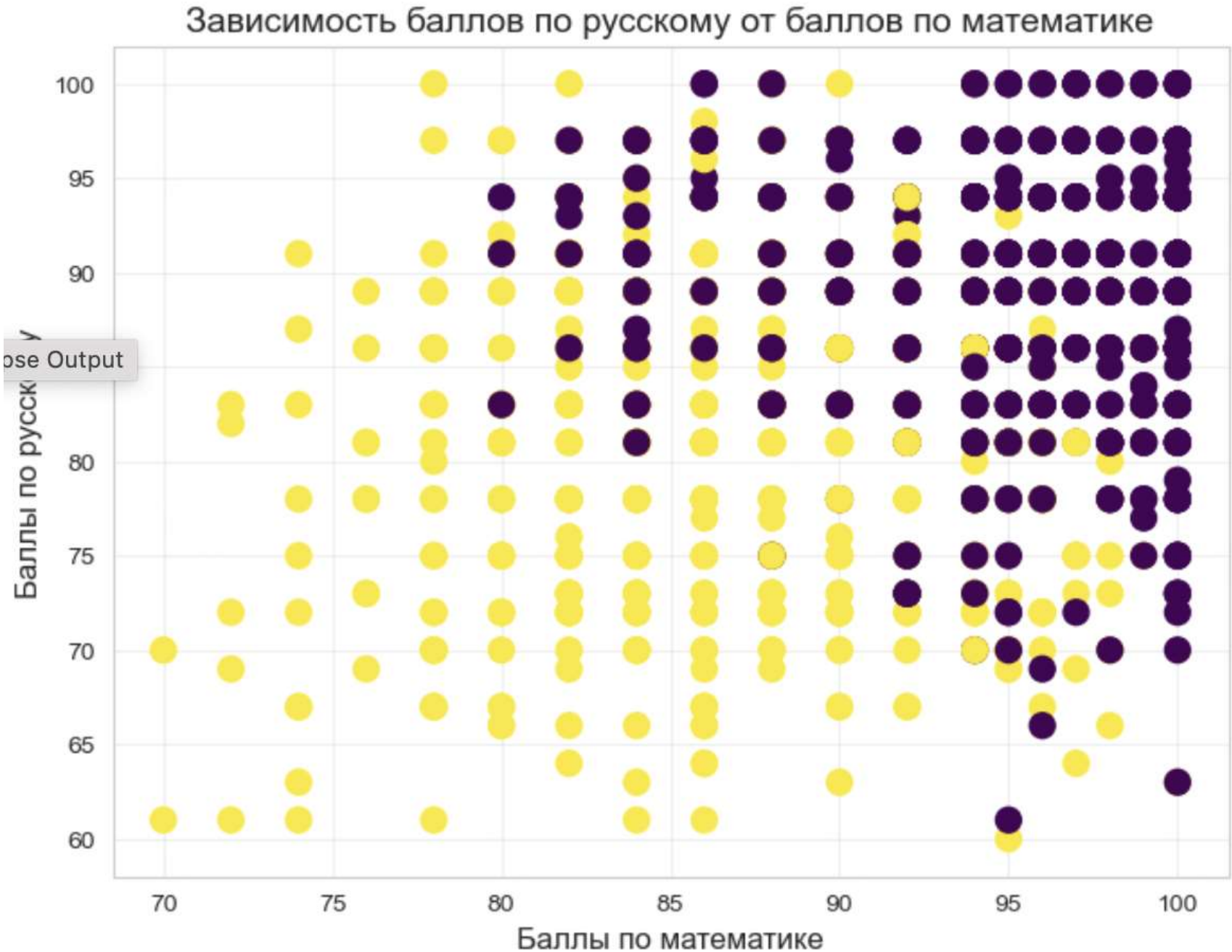
Есть ли подобные корелляции в случае с парой "Физика"/"Математика"?

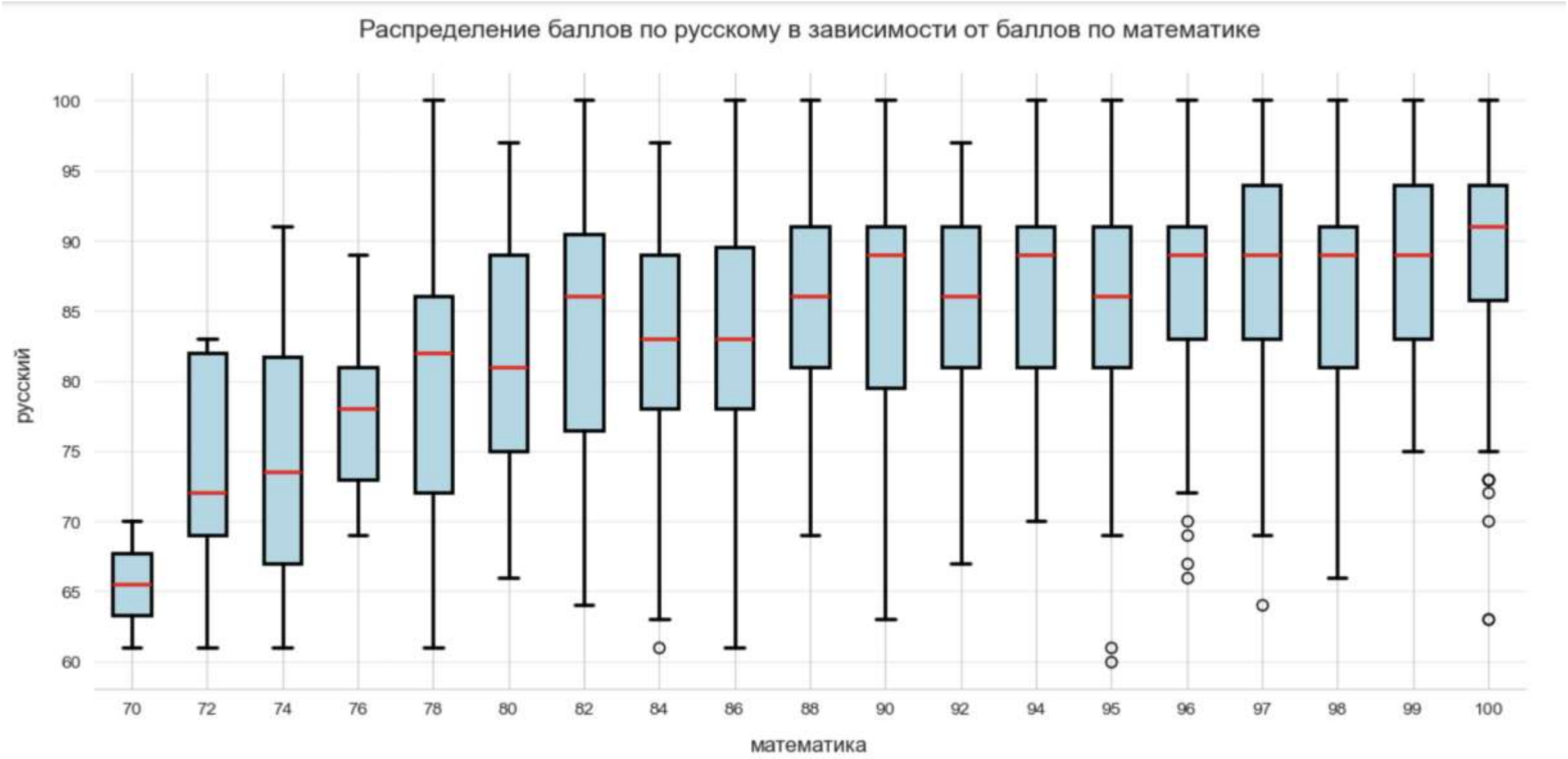



Распределение баллов по физике в зависимости от баллов ДВИ
Баллы по физике
100
90
80
70
60
50
55
60
65
70
75
80
85
90
95
100
Баллы ДВИ


Распределение баллов по физике (без выбросов)
физика
100
90
80
70
60
50
55
60
65
70
75
80
85
90
95
100
ДВИ


Зависимость баллов по физике от баллов ДВИ
Баллы по физике
100
90
80
70
60
50
60
70
80
90
100
Баллы ДВИ


Зависимость баллов по русскому от баллов по математике
100
95
90
80
75
70
65
60
70
75
80
85
90
95
100
Баллы по математике



С русским языком такой корелляции не прослеживается. Это объяснимо: знание физики и математики связано друг с другом, как и способность студента одновременно хорошо сдать математику ДВИ и математику ЕГЭ. С русским языком же такой зависимости не наблюдается.

- В целом, корреляция между баллами по математике и по русскому прослеживается, но она не выражена явно.
- Большой разброс баллов по русскому среди студентов с низкими и средними баллами по математике может указывать на то, что подготовка по русскому среди этой группы более разнородная.
- В группе с высокими баллами по математике разброс по русскому уменьшается, но остаются единичные выбросы, которые требуют дополнительного анализа (например, причин низких баллов по русскому у сильных математиков).

Это объяснимо: знание физики и математики связано друг с другом, как и способность студента одновременно хорошо сдать математику ДВИ и математику ЕГЭ. С русским языком же такой зависимости не наблюдается.