Федеральное государственное образовательное бюджетное учреждение

высшего образования

«Финансовый университет при Правительстве Российской Федерации»

Департамент анализа данных, принятия решений и финансовых технологий

Курсовая работа

по дисциплине «Теория вероятностей и математическая статистика»

на тему:

«Проверка гипотезы о нормальном распределении

логарифмической доходности по критериям Хегази-Грина»

Вид исследуемых данных:

«Котировки акций компаний, входящих в индекс ММВБ финансы»

Выполнил:

студент группы ПМ19-1

Быханов Н.Ю.

Научный руководитель:

профессор ДАДиМО,

д.э.н Коровин Дмитрий Игоревич

Москва 2021

Оглавление

Введение**3**

I. Предварительный анализ данных 5

II. Теоретическая справка11

1.Математическая статистика……………………………………………………….....…11

2.Статистическая гипотеза**11**

3.Ошибки первого и второго рода…………………………………………………….….12

6.P-значение (P-value) статистического критерия13

7.Критерии Хегази-Грина....................................................................................................13

8.Логарифмическая доходность……………………………………………….……….....14

III. Практическая часть ........................................................................................................15

1. Выбор альтернативной гипотезы и оценка мощности критерия ................................15

2.Проверка гипотез для реальных данных.........................................................................15

Заключение .........................................................................................................................24

Список используемых источников....................................................................................25

Приложения ....................................................................................................................... 25

Приложение 1(Модельные данные)………………………………………………………25

Приложение 2…………………………………................................................................ 25

Технические характеристики компьютера ………..........................................................25

Время выполнения программ …………………….......................................................... 21

Приложение 2 ………………........................................................................................ 26

Коды программ ………………....................................................................................... 26

Приложение 3 ……………….......................................................................................... 51

**ВВЕДЕНИЕ**

В данной работе первым этапом ставим гипотезу о логарифмической доходности акций компаний входящих в индекс ММВБ Финансы, распределенных по нормальному закону. На втором этапе обработки проверяем гипотезу с помощью методов математической статистики с помощью критерия Хегази-Грина.

Отмечаем целью данной работы – выяснить и соотнести с реальностью гипотезу о нормальном распределении логарифмической доходности по выбранному критерию Хегази-Грина. На практике  критерий Хегази-Грина  используется крайне редко, я постараюсь доказать что данный критерий прост и эффективен. Работа выполнена в несколько этапов: для начала следует убедиться в правильности критерия,  для этого проведем проверку гипотезы на модельных данных. Затем рассмотрим альтернативную гипотезу, выберем из всех предложенных альтернатив самую близкую к основной гипотезы и только после этого  дадим оценку мощности критерия Хегази-Грина.

Котировки выбранного нами индекса будут рассматриваться есть за период с 01 января 2014 года по 01 января 2021 года.

Мною были взяты несколько разных таймфреймов дневной и часовой на период от начала санкций и присоединения Крыма и до пандемии - с 01 января 2014 года по 1 января 2018 года и также дневной для промежутка, когда была пандемия 1 января 2018 по 1 января 2021 года. Данные часовые будут взяты квартально в середине нашего периода, 2016 год, а то есть 1 января 2016 – 31 марта 2016, как тогда уже санкции прошли и акции вели себя нормально и без вмешивания в ценообразование политики.

Индекс ММВБ является ценовым композитным фоновым индексом, который взвешен по рыночной капитализации. Он включает в себя акции крупнейших российских эмитентов, которые являются наиболее ликвидными и активно развивающимися.  А также на этих акциях создано большинство торговых ботов. Для лучшего обучения и использования нейросети нужно подавать на вход нормально распределённые данные. В курсовой работе мы будем проверять, является ли всегда логарифмическая доходность, выбранных нами котировок, нормально распределенной. Потому что эта компонента часто используется на вход в нейросеть.

Индекс ММВБ Финансы изучает акции компаний, чья основная деятельность заключается в оказании банковских и финансовых услуг. Например, вклады, кредитование, обслуживание банковских карт и так далее.  Исследуемый индекс на апрель 2020 года состоит из акций 8 компаний, таких как БСП (ПАО «Банк Санкт-Петербург»), ОАО Московская биржа , ОАО Московский Кредитный банк, QIWI, ПАО САФМАР Финансовые инвестиции, ПАО Сбербанк России, ПАО банк ВТБ. Возможно сокращение количества рассматриваемых компаний в связи с недостаточным объёмом данных. Например – малое число торговых дней.

Работа состоит из двух этапов: обзор теоретических аспектов и на предварительный анализ данных, который поможет исследовать мощность критерия.  После получения результатов, получим выводы, сделанные на проверки гипотезы на модельных и реальных данных по приоритетному первому критерию Хегази-Грина.

Актуальность работы заключается в использовании современного инструмента обработки данных, а именно языка программирования Python, который позволяет быстро обрабатывать данные и имеет большое количество библиотек для анализа и обработки данных. Кроме того, сфера банковских и финансовых услуг, к которой относится исследуемый индекс, в 2021 году пользуется большим спросом.

Проверка гипотезы на модельных и эмпирических данных будет осуществляться при помощи графической веб-оболочки Jupyter notebook для написания воспроизводимого кода Python, посредством которого будут проведены все расчеты, а также построение графических объектов и таблиц.

**I. ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ**

Проведём первичный анализ данных. Информация о котировках акций была взяла с официального сайта Финам: fiman.ru. Данные об индексе, и входящих в него акциях с сайта <https://www.investing.com/indices/micex-financials-components>. В таблице 1 представлено сопоставление названий тикеров и компаний.

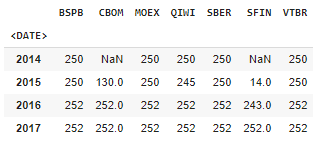
Таблица 1 - Список компаний и тикеров котировок акций

|  |  |
| --- | --- |
| Тикер | Название компаний |
| BSPB | ПАО «Банк Санкт-Петербург», |
| MOEX | ОАО Московская биржа |
| CBOM | ОАО Московский Кредитный банк |
| QIWI | QIWI |
| SFIN | ПАО САФМАР Финансовые инвестиции |
| VTBR | ПАО банк ВТБ |
| SBER | ПАО Сбербанк России |

Начнем для периода с 2014-2018 года с таймфреймом день.

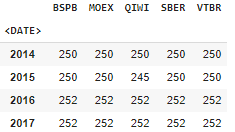
Проведем расчет количества торговых дней с последующим выводом результатов в таблицу, где по столбцам названия компании, а по строчкам – года. Для этого будет использовать поле DATE, показывающее дату информации о котировках. Оно представлено в формате дд/мм/гг. Для вычисления торговых дней используется группировка данных по полю DATE, используя лишь данные о годе, после чего осуществляется подсчёт торговых дней в получившихся группах и дальнейший вывод получившегося результата в таблицу с дополнительным экспортом в файл формата csv с разделителем «;».

Полученные результаты количества торговых дней для всех компаний можем наблюдать в таблице 2.[3]



Отметим, что у котировок CBOM, SFIN данные есть лишь с 15 года, , поэтому мы не анализируем их в дальнейшем. Составим новую таблицу дней, где будут только те компании, в которых есть торговые дни в период с 2014 по 2018 года. Полученные результаты представлены в таблице 3.

Таблица 3 - Вычисленное количество торговых дней для сокращенного списка компаний.[4]



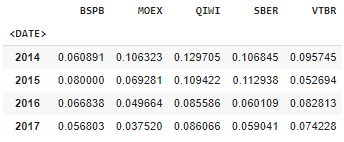
Оставшиеся 4 компании имеют 248 и более торговых дней в каждом из рассматриваемых годов, что обеспечивает достаточное количество информации для дальнейшего анализа и обработки.

Далее рассмотрим максимальные относительные скачки цен вверх и вниз. Нам потребуется поле <CLOSE>, показывающее цену последней сделки, которая была совершена. В таблице 3 мы можем наблюдать полученный результат для максимальных скачков цен вниз, а в таблице 4 – для максимальных скачков вверх.

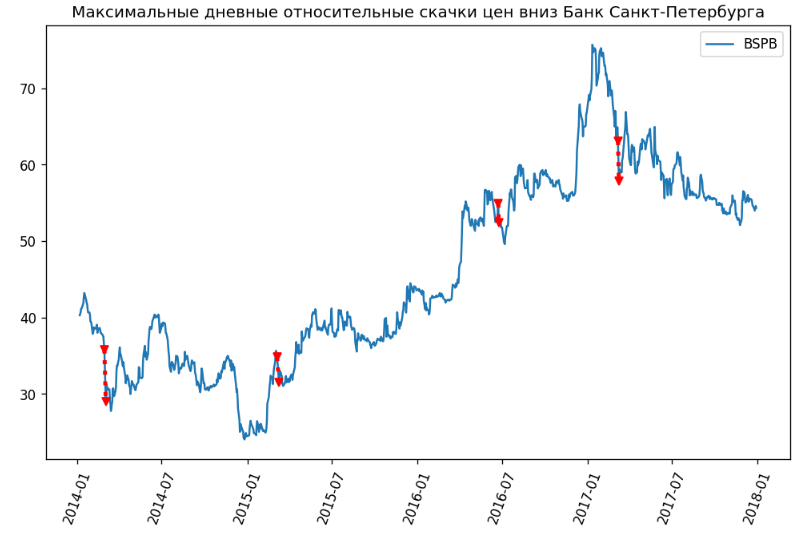
Таблица 4 - Вниз максимальные относительные скачки цен[5]



Таблица 5 - Вверх максимальные относительные скачки цен.[5]



Максимальный скачок вниз у BSPB (23%) и QIWI сильный скачок вверх (12,9%), но и далеко не отошли акции BSPB (8%) благодаря чему можно прийти к выводу о том, что обладает максимальными относительными изменениями цен. Построим график изменения цен для этой компании. Результаты отображены на рисунках 1 и 2, на котором показаны сильные скачки акций. Можно заметить, что очень сильные максимальные движения вниз в 2014 году, это связанно из-за политической ситуации и начала вступления санкций на Россию, что дало очень сильные изменения в цене.



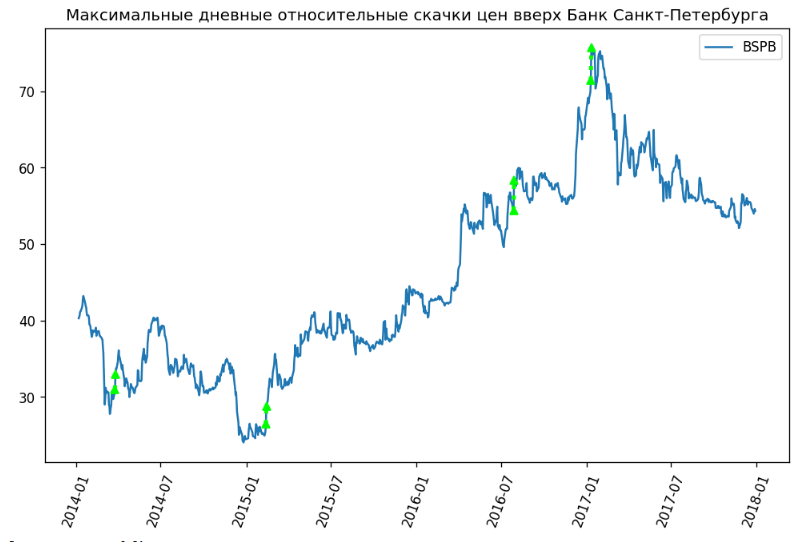
Рисунок 1. График изменения цен для компании BSPB максимальные скачки вниз.[11]

Рисунок 2. График изменения цен для компании BSPB максимальные скачки вверх.[11]

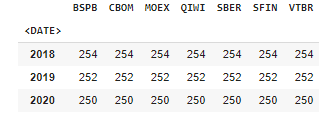
Остальные котировки оставляем, как скачки цен оказались незначительными, а значит можно смело приступать к практической части работы с данными периода 2014-2018.

Следующий период с 2018-2020 года с таймфреймом день.

Уже для нового набора данных мы начинаем наш анализ.

Посчитаем количество торговых дней для данного датасета.

Таблица 6 - Результаты количества торговых дней для всех компаний.[6]



Как уже у всех акций есть достаточное количество дней, то мы не убираем из анализа акции компаний.

Далее рассмотрим максимальные относительные скачки цен вверх и вниз, как и для прошлого датасета.

Таблица 7 - Вверх максимальные относительные скачки цен. [9]

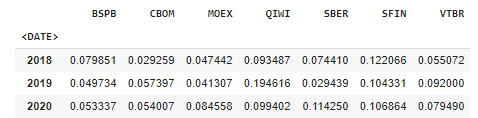


Таблица 8 - Вниз максимальные относительные скачки цен.[10]



Максимальный скачок вниз у SFIN (35%) и сильный скачок вверх (12%), благодаря чему можно прийти к выводу о том, что обладает максимальными относительными изменениями цен. Построим график изменения цен для этой компании. Результаты отображены на рисунках 1 и 2, на котором показаны сильные скачки акций. Можно заметить, что очень сильные максимальные движения вниз в 2018 году, это связанно из-за рода деятельности компании, они занимаются инвестициями и страхованием, а в 2018 году была большая смертность и начались страховые случаи, которые сильно задели по благонадежности компании и ее выручке.

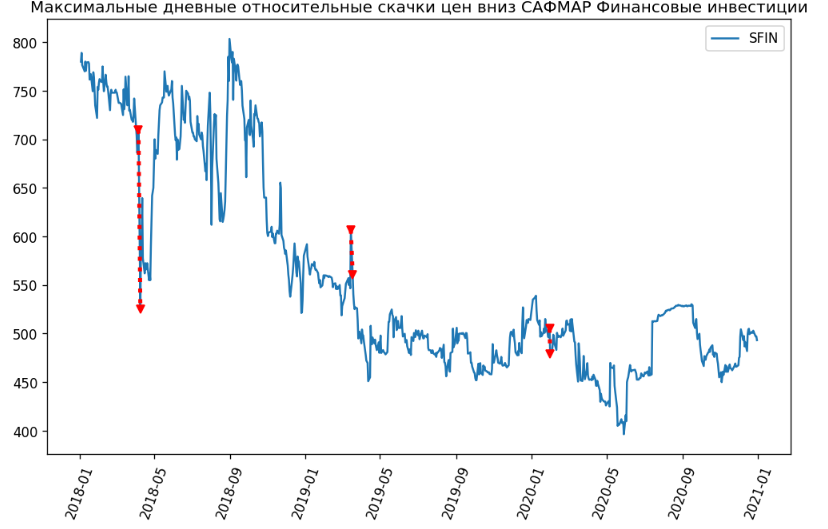


Рисунок 3. График изменения цен для компании SFIN максимальные скачки вниз[11]

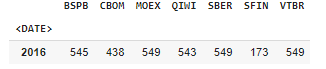


Рисунок 4. График изменения цен для компании SFIN максимальные скачки вверх[11]

Все компании пригодны для качественной проверки гипотезы о нормальном распределении логарифмической доходности по выбранному критерию, ни один тикер не был исключен, ведь даже скачки цен оказались незначительными, а значит можно смело приступать к практической части работы с периодом 2018 -2020.

Следующий период первый квартал 2016 года с таймфреймом час.

Таблица 9 - Вычисленное количество торговых дней для сокращенного списка компаний. [11]



Можно заметить, что у SFIN имеет слишком мало было торговых часов, поэтому мы в будущем не анализируем.

**II. ТЕОРЕТИЧЕСКАЯ СПРАВКА**

**Математическая статистика**

Математическая статистика является частью единой дисциплины прикладной математики, которую называют теорией вероятностей и математической статистикой и представляет из себя совокупность или же семейство дисциплин, таких как экономическая статистика, финансовая статистика, социальная статистика и так далее, основной задачей которой является обеспечение конкретных статистических дисциплин теоретическим фундаментом.

Цель математической статистики можно обозначить как создание методов сбора и обработки статистических данных для дальнейшего получения как практических, так и научных выводов.

Статистическая гипотеза

Статистической гипотезой можно назвать всякое утверждение, содержащее информацию о виде или параметрах внутреннего распределения, об отношениях между случайными величинами и тому подобное. При этом, её можно считать параметрической, если она базируется на предположении, заключающемся в том, что нам известно генеральное распределение, причём с точностью до конечного числа параметров.

Пусть H0 и H1 – это две статистические гипотезы, являющиеся взаимоисключающими, при этом назовём гипотезу H0 основной, а H1 – вспомогательной. В дальнейшем принимаем в качестве базисного предположения утверждение о том, что одна из гипотез является справедливой. Чтобы проверить нулевую гипотезу, применяют случайную величину, специально подобранную, для которой известно точное или приближённое распределение, и называют данную величину статистическим критерием или статистикой критерия. Совокупность значений данного критерия, при которых принимается нулевая гипотеза, называют областью допустимых значений или же областью принятия гипотезы. В противном случае, то есть те значения статистического критерия, в которых гипотеза отвергается, область называется критической.

Как правило критическую область можно задать, используя неравенства:

(1.1)

или

(1.2)

или

, (1.3)

где , - статистика критерия.

**Ошибки первого и второго рода**

Во время применения статистики критерия возможно возникновение ошибок двух различных типов. Случай, когда отвергается гипотеза H0, являющаяся верной, называют ошибкой первого рода. Напротив, ошибка второго рода – ситуация, при которой отвергается гипотеза H1, являющаяся верной.

Обозначим за α вероятность ошибки первого рода, которая называется значимостью критерия, а β – вероятность ошибки второго рода, при этом величина 1-β называют мощностью критерия.

Статистическая гипотеза проверяется путем сравнения наблюдаемого значения критерия с критическим значением, связанным с данным уровнем значимости, что позволяет отклонить или принять основную гипотезу. При этом в тех случаях, когда уровень значимости будет другим, то придётся вновь

**P-значение (P-value) статистического критерия**

Данное понятие стало распространённым из-за обширного применения статистических и оно обеспечивает решение вопроса, связанным с принятием или отклонением гипотезы, являющейся основной, при этом выполняет это одновременно для всех уровней значимости, и нет необходимости вычислять критические значения.

Р-значением статистического критерия для фиксированной реализации x случайной называется такое число PV(x)  , что  для любого уровня значимости α, при котором гипотеза H0 принимается, и , для любого уровня значимости α, при котором гипотеза H0 отвергается.

Предположим, что Р-значение PV(x) уже найдено или предварительно известно, тогда решение о принятии или отклонении гипотезы H0 для заданного α осуществляется на основе следующего простого правила: в случае, когда , гипотеза H0 отвергается, а если гипотеза H0 принимается.

Проверка гипотезы при помощи P-значения более информативна, нежели традиционная проверка с помощью критического значения. Р - значение с гораздо большей точностью, чем обычные способы проверки статистических гипотез. Тем не менее, выбор того или иного способа проверки зависит от наличия соответствующих таблиц или компьютерных программ. При верной основной гипотезе P-значение равномерно распределено на отрезке [0,1]. Поэтому вероятность получить малое равна вероятности получить большое P-значение . Однако, если H0 не верна, наблюдаемые P-значения (при достаточно высокой мощности критерия) концентрируются около нуля.

Р - значение находится из равенства: .

Вспомогательный критерий для проверки статистической гипотезы, Критерий Колмогорова.

Обозначим как то количество составляющих вектора , которые меньше х для любого . Если мы имеем дело со случайным вектором , то обозначение аналогично, хотя стоит отметить, что оно будет являться дискретной случайной величиной, которая может принимать целые значения от 0 до n. Обозначим за x реализацию случайной выборки X, имеющую объём n, из некоторого распределения с функцией F(x), и тогда эмпирическая функция распределения, которая соответствует данной выборке будет иметь вид:

. (4.1)

Для оценки функции F(x) по случайной выборке X формула аналогична и выглядит следующим образом:

. (4.2)

Отметим, что является числовой функцией, в то время как является случайным процессом, поскольку в каждой точке x принимает случайное значение.

Расстояние между функциями и F(x) можем определить, используя формулу:

. (4.3)

Аналогично обратим внимание на то, что в случае с функцией функции расстояние будет являться числом, но для расстояние уже случайная величина а значит принимает значения от 0 до 1 включительно.

Опираясь на теоремы Колмогорова в случае непрерывной функции F(x) при любом неотрицательном u0 существует предел

, (4.4)

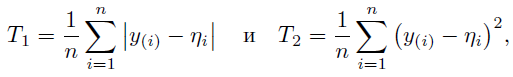
где

(4.5)

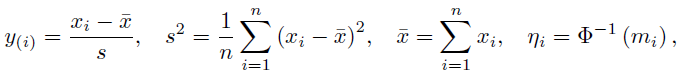
А также на основе данной теоремы критерий согласия с критической областью , где - корень уравнения K(u) = 1-а, где уровень значимости стремится к при , то есть является асимптотическим уровнем значимости. Этот критерий и является критерием Колмогорова, и применяется при n> 20, так как в противном случае фактический уровень значимости будет сильно отличаться от номинального значения.

Критерии Хегази-Грина

В работе в качестве основного статистического критерия применяются критерии Хегази-Грина, которые относится к группе критериев, где рассматриваются отклонения статистики порядка от их математических ожиданий (по медианам и др.). К достоинствам этого критериев можно отнести мощность относительно других, например, критерия Шапиро-Уилка, Андерсона-Дарлинга, Д'Агостино. Статистики критерия Хегази-Грина:



где



Ф – функция распределения N(0;1), mi – количество элементов выборки,

– вероятность попасть в i-й элемент разбиения.

Чтобы вычислить критерий, нужно проанализировать значения процентных точек для статистик T1 и T2. Проверяемая гипотеза отклоняется при больших значениях статистик, также важно чтобы в статистике использовалась несмещенная оценка дисперсии. Подчеркнем, что в статистике должна использоваться именно несмещенная оценка дисперсии.

**Логарифмическая доходность**

Логарифмическую доходность или данные об изменение цен довольно часто используют для анализа данных. Данный показатель используется вместо процентной доходности и имеет вид:

(3.1)

Или же, применив свойство логарифма от частного, получим:

(3.2)

Где t – рассматриваемый период, – цена акции в данный период, – цена акции за прошлый период.

Стоит отметить ряд достоинств логарифмической доходности. Во-первых, она позволяет без особых проблем объединять доходы при более низких частотах выплат, что возможно только засчёт обобщения доходов при более высоких частотах выплат, а значит, ежемесячная доходность будет равна сумме ежедневных логарифмических доходностей, что делает расчёты проще, нежели с использованием процентной доходности. Также логарифмическая доходность позволяет найти начальную цену, если произошло её повышение на x%, а затем снижение на такую же величину x%.

**III. ПРАКТИЧЕСКАЯ ЧАСТЬ**

1.Выбор альтернативной гипотезы и оценка мощности критерия

Выбирая альтернативную гипотезу, следует учитывать, что она должна быть близка к нулевой при некоторых значениях или вариантах проверяемых данных. Для данной работы будет уместно взять в качестве альтернативной гипотезы предположение о том, что логарифмическая доходность имеет распределение Стьюдента, так как оно сводится к нормальному при больших значениях статистик, для нашей проверки возьмем со степенями свободы 8. С помощью программы «Мощность критерия» оценим мощность разных уровней значимости критерия Хегази-Грина (T1) и (T2) для выбранного распределения.

Таблица 10. Мощность критерия Хегази-Грина[14]

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1% | 5% | 15% |
| T1 | 0,97 | 0,85 | 0,68 |
| T2 | 0,97 | 0,88 | 0,73 |

Из таблицы видно, что мощность достаточно высокая при различных доверительных интервалах, значит ошибка второго рода маловероятна, критерий Хегази-Грина имеет высокую мощность.

**2.  Проверка гипотезы для реальных данных**

Проверив работоспособность алгоритма, можно приступить к проверке гипотезы о нормальном распределении на реальных данных. Используем программы «Проверка на реальных данных Т1» (применяя статистику Хегази – Грина 1) и «Проверка на реальных данных Т2» (применяя статистику Хегази – Грина 2) для нахождения P-значения по всем годам для всех компаний, которые мы используем в соответствии наших выводов по предварительному анализу данных с период 2014 – 2018 года с таймфреймом день. То есть мы берем из сокращенного списка компаний: BSPB,MOEX,QIWI,SBER,VTBR :

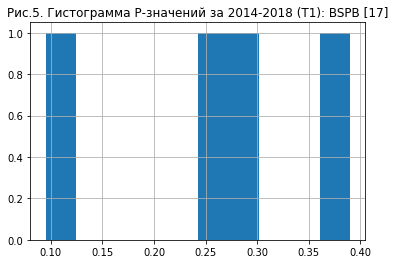
Таблица 11. P-значения для реальных данных (T1)[15]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Тикер** | **2014** | **2015** | **2016** | **2017** |
| BSPB | 0,095 | 0,394 | 0,295 | 0,263 |
| MOEX | 0,383 | 0,462 | 0,298 | 0,263 |
| QIWI | 0,029 | 0,612 | 0,088 | 0,899 |
| SBER | 0,081 | 0,116 | 0,313 | 0,04 |
| VTBR | 0 | 0,005 | 0,025 | 0,156 |

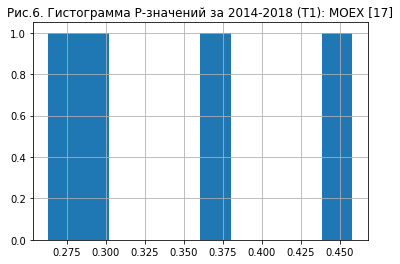
Таблица 12. P-значения для реальных данных (T2)[16]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Тикер | 2014 | 2015 | 2016 | 2017 |
| BSPB | 0,618 | 0,641 | 0,109 | 0,285 |
| MOEX | 0,341 | 0,505 | 0,351 | 0,334 |
| QIWI | 0,206 | 0,552 | 0,067 | 0,686 |
| SBER | 0,59 | 0,458 | 0,186 | 0,008 |
| VTBR | 0 | 0,095 | 0,038 | 0,527 |

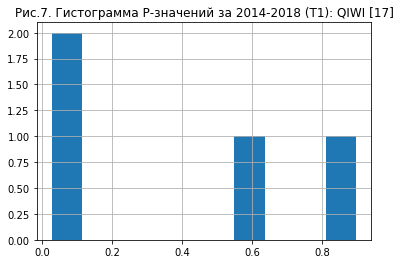
При визуализации данных значений получаем следующие диаграммы:



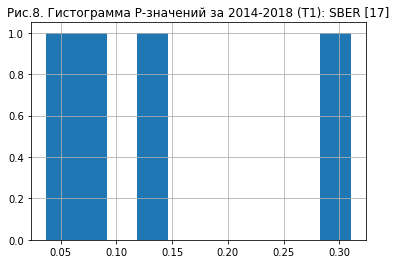
Можно сделать вывод, что Р-значения у котировок BSPB неравномерно распределены на всем отрезке, что подтверждается Р-значением критерия Колмогорова, которое получилось равным 0.04000.



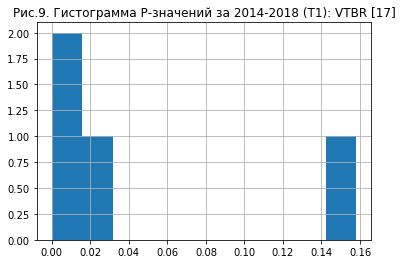
Можно сделать вывод, что Р-значения у котировок MOEX равномерно распределены на всем отрезке, что подтверждается Р-значением критерия Колмогорова, которое получилось равным 0.12702.



Можно сделать вывод, что Р-значения у котировок MOEX равномерно распределены на всем отрезке, что подтверждается Р-значением критерия Колмогорова, которое получилось равным 0.4702.



Можно сделать вывод, что Р-значения у котировок SBER неравномерно распределены на всем отрезке, что подтверждается Р-значением критерия Колмогорова, которое получилось равным 0.019.



Можно сделать вывод, что Р-значения у котировок VTBR неравномерно распределены на всем отрезке, что подтверждается Р-значением критерия Колмогорова, которое получилось равным 0.0012.

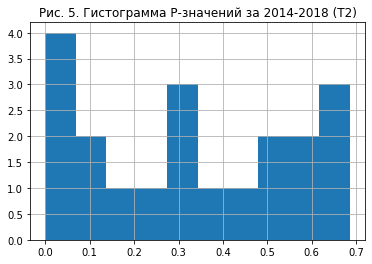
****

Рисунок 10 – Гистограмма Р-значений за 2014-2018 (T2) для вех компаний сразу[18]

По данным гистограммам заметно, что P-значения распределены не равномерно у большинства графиков, что и подтверждает критерий Колмогорова. Таким образом, равномерность отвергается у котировок Сбера и ВТБ, у других либо маленький P-value или очень не высокий.

Вычислим долю проверок, в которых гипотеза принималась при 5% и 1% уровнях значимости (по Т1). Голубым отмечены случаи принятия гипотезы при уровне значимости в 1%, оранжевым – случаи принятия гипотезы при уровнях значимости 1% и 5%. Таким образом, при **α** = 5 % гипотеза отвергалась в 25% случаях, а при **α**= 1 % в 10% случаев. Из этого следует, что предполагаемая гипотеза отвергается.

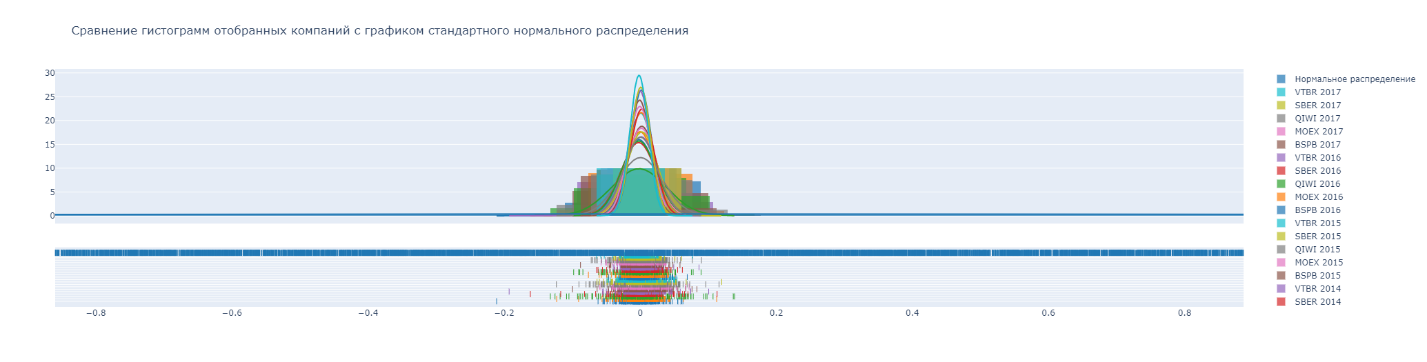
****

Рисунок 11 - Сравнение гистограмм отобранных компаний с графиком стандартного нормального распределения в случае полного года[19]

Из гистограммы следует, что выбранные периоды немного схожи с нормальным распределением, выделенным синим цветом, хотя имеют не свойственные ему значения. Логдоходность «ПАО банк ВТБ» сосредоточена в отрезке [-0.2;0.2], а в основном все компании находятся на этом промежутке также сильно вытянувшись, подробнее можно посмотреть в Jupyter Notebook пользуясь преимуществами интерактивных графиков библиотеки plotly.

В случае годовых значений гипотеза о нормальности отвергается с помощью графического метода, но доля тех данных, которые распределены по нормальному закону, хоть и является меньшей частью, но уже вполне значительна и должна быть принята во внимание.

Это можно объяснить тем, что в этот период были первые санкции, но наши компании не были в первых списках, поэтому у них наблюдается тенденция к нормальному распределению, что видно на многочисленных графиках, но из-за хаоса на рынке нормальности данных не наблюдается. Что и подтверждают два критерия Хегази- Грина.

Используем дальше программы «Проверка на реальных данных Т1 для часа» (применяя статистику Хегази – Грина 1) и «Проверка на реальных данных Т2 для часа» (применяя статистику Хегази – Грина 2) для нахождения P-значения по всем годам для всех компаний, которые мы используем в соответствии наших выводов по предварительному анализу данных, сужая интервал исследования, с период 2018 – 2020 года с таймфреймом день. Из-за того, что пандемия вызывала высокий уровень волатильности и хаоса на рынке ценных бумаг.

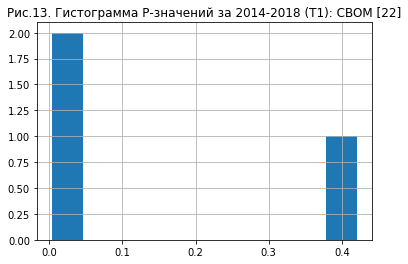
Проанализируем такой список компаний: BSPB,CBOM,MOEX,QIWI,SBER,SFIN,VTBR

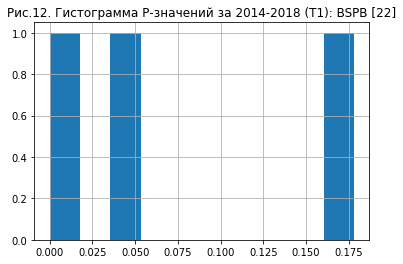
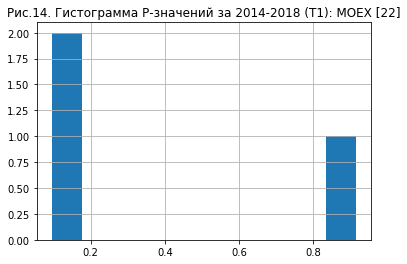
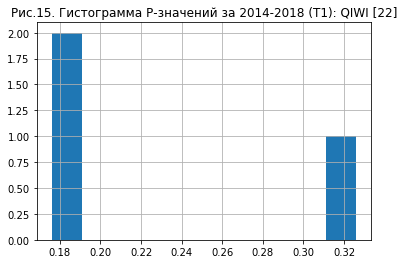
Таблица 13. P-значения для реальных данных (T1) [20]

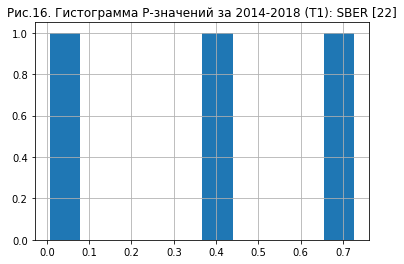
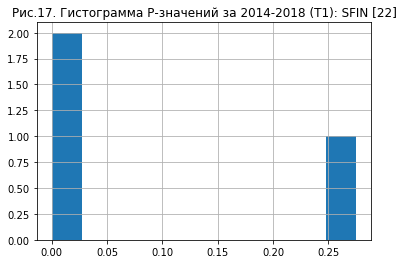
|  |  |  |  |
| --- | --- | --- | --- |
| **Тикер** | **2018** | **2019** | **2020** |
| BSPB | 0,042 | 0,179 | 0 |
| CBOM | 0,423 | 0,004 | 0,028 |
| MOEX | 0,914 | 0,1 | 0,095 |
| QIWI | 0,189 | 0,177 | 0,331 |
| SBER | 0,411 | 0,728 | 0,008 |
| SFIN | 0,275 | 0,023 | 0 |
| VTBR | 0,717 | 0,057 | 0,028 |

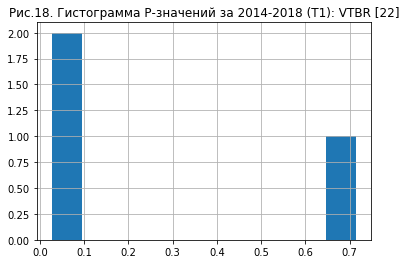
Таблица 14. P-значения для реальных данных (T2)[21]

|  |  |  |  |
| --- | --- | --- | --- |
| **Тикер** | **2018** | **2019** | **2020** |
| BSPB | 0,676 | 0,292 | 0,022 |
| CBOM | 0,174 | 0,765 | 0,057 |
| MOEX | 0,851 | 0,241 | 0,202 |
| QIWI | 0,293 | 0,581 | 0,75 |
| SBER | 0,474 | 0,554 | 0,031 |
| SFIN | 0,301 | 0,047 | 0,099 |
| VTBR | 0,782 | 0,402 | 0,065 |

При визуализации данных значений получаем следующие диаграммы:







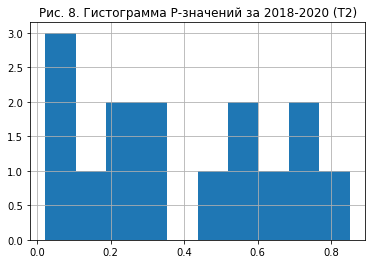
****

Рисунок 19 – Гистограмма Р-значений за 2018-2020 (T2) )[22]

По данным гистограммам заметно, что P-значения распределены не равномерно. Что и подтверждает критерий Колмогорова на всех гистограммах, кроме гистограммы SBER с p-value 0.90.

Вычислим долю проверок, в которых гипотеза принималась при 5% и 1% уровнях значимости (по Т1). Голубым отмечены случаи принятия гипотезы при уровне значимости в 1%, оранжевым – случаи принятия гипотезы при уровнях значимости 1% и 5%. Таким обратим, при **α** = 5 % гипотеза отвергалась в 38% случаях, а при **α**= 1 % в 19% случаев. Из этого следует, что предполагаемая гипотеза отвергается. И можно заметить, что во время коронавируса – процент принятия гипотезы стал намного ниже.

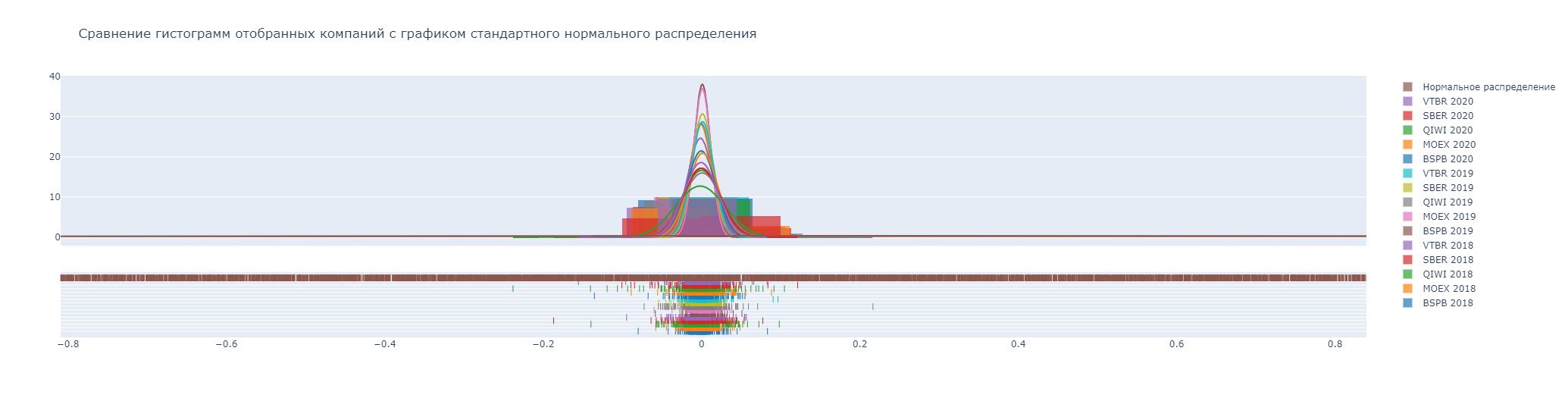
****

Рисунок 10 - Сравнение гистограмм отобранных компаний с графиком стандартного нормального распределения в случае полного года[23]

Из гистограммы следует, что выбранные периоды немного схожи с нормальным распределением, выделенным синим цветом, хотя имеют не свойственные ему значения. Логдоходность ПАО «Банк Санкт-Петербург», сосредоточена в отрезке [-0.1;0.1], а в основном все компании находятся на этом промежутке также сильно вытянувшись, подробнее можно посмотреть в Jupyter Notebook пользуясь преимуществами интерактивных графиков библиотеки plotly.

В случае годовых значений гипотеза о нормальности отвергается с помощью графического метода, но доля тех данных, которые распределены по нормальному закону, хоть и является меньшей частью, но уже вполне значительна и должна быть принята во внимание.

Это можно объяснить тем, что в этот период был период начала пандемии, поэтому тоже была неопределенность на рынке и наши компании напрямую связаны с финансовым положением россиян, как котировки это либо банки, либо предлагающие финансовые услуги. Поэтому резкое падение уровня доходов россиян и большая волатильность могла так сказаться на наших данных.

Также давайте проверим логдоходность на нормальность с помощью статистики Колмогорова-Смирнова на этом же интервале.

|  |  |  |  |
| --- | --- | --- | --- |
| **Тикер** | **2018** | **2019** | **2020** |
| BSPB | 0,060638 | 0,006932 | 0,073101 |
| CBOM | 0,178377 | 2,77E-10 | 0,446313 |
| MOEX | 0,000412 | 0,185557 | 3,9E-06 |
| QIWI | 3,24E-08 | 3,64E-06 | 3,9E-07 |
| SBER | 4,22E-06 | 0,00723 | 2,32E-06 |
| SFIN | 4,98E-10 | 0,00685 | 0,027148 |
| VTBR | 0,018033 | 0,549497 | 0,013517 |

Таблица 14. P-значения для реальных данных Колмогоров – Смирнов [29]

Вычислим долю проверок, в которых гипотеза принималась при 5% и 1% уровнях значимости по статистике Колмогорова -Смирнова. Голубым отмечены случаи принятия гипотезы при уровне значимости в 1%, оранжевым – случаи принятия гипотезы при уровнях значимости 1% и 5%. Таким обратим, при **α** = 5 % гипотеза отвергалась в 70% случаях, а при **α**= 1 % в 52% случаев. Из этого следует, что предполагаемая гипотеза отвергается. И можно заметить, что во время коронавируса – процент принятия гипотезы стал намного ниже.

Частота принятия H0 уменьшилась по сравнению с первой статистикой Хегази-Грина, по крайней мере в 2 – 3 раз. Критерий Колмогорова-Смирнова показал большую частоту отвержения гипотезы. Быть может, именно поэтому данный критерий не используют как основной для исследования на нормальность, ведь это общий критерий согласия, более универсальный.

Используем дальше программы «Проверка на реальных данных Т1 для часа» (применяя статистику Хегази – Грина 1 для часа) и «Проверка на реальных данных Т2 для часа» (применяя статистику Хегази – Грина 2) для нахождения P-значения по всем годам для всех компаний, которые мы используем в соответствии наших выводов по предварительному анализу данных с первый квартал 2016 года с таймфреймом час.

То есть мы берем списка компаний: BSPB,CBOM,MOEX,QIWI,SBER,VTBR.

Таблица 15. P-значения для реальных данных (T1) [24]

|  |  |
| --- | --- |
| **Тикер** | **2016** |
| BSPB | 0 |
| CBOM | 0,074 |
| MOEX | 0,041 |
| QIWI | 0,022 |
| SBER | 0,05 |
| VTBR | 0,092 |

Таблица 16. P-значения для реальных данных (T2) [26]

|  |  |
| --- | --- |
| **Тикер** | **2016** |
| BSPB | 0,094 |
| CBOM | 0,29 |
| MOEX | 0,291 |
| QIWI | 0,111 |
| SBER | 0,101 |
| VTBR | 0,06 |

При визуализации данных значений получаем следующие диаграммы:

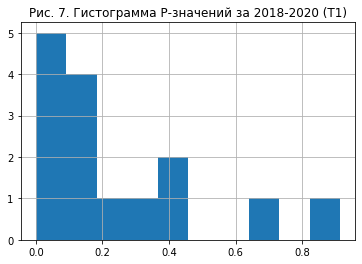


Рисунок 11 – Гистограмма Р-значений за 2016-2016 (T1) [25]

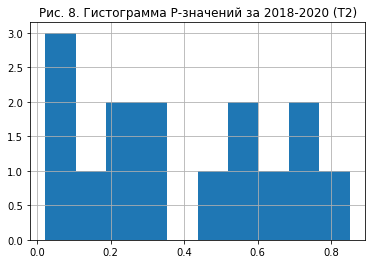
****

Рисунок 12 – Гистограмма Р-значений за 2016-2016 (T2) [27]

По данным гистограммам заметно, что P-значения распределены не равномерно. Таким образом, равномерность отвергается. Что и показывает Критерий Колмогорова, показывая на каждом очень низкое значение p-value ниже 0.05 и 0.01.

Вычислим долю проверок, в которых гипотеза принималась при 5% и 1% уровнях значимости (по Т1). Голубым отмечены случаи принятия гипотезы при уровне значимости в 1%, оранжевым – случаи принятия гипотезы при уровнях значимости 1% и 5%. Таким образом, при **α** = 5 % гипотеза отвергалась в 50% случаях, а при **α**= 1 % в 16% случаев. Из этого следует, что предполагаемая гипотеза отвергается.

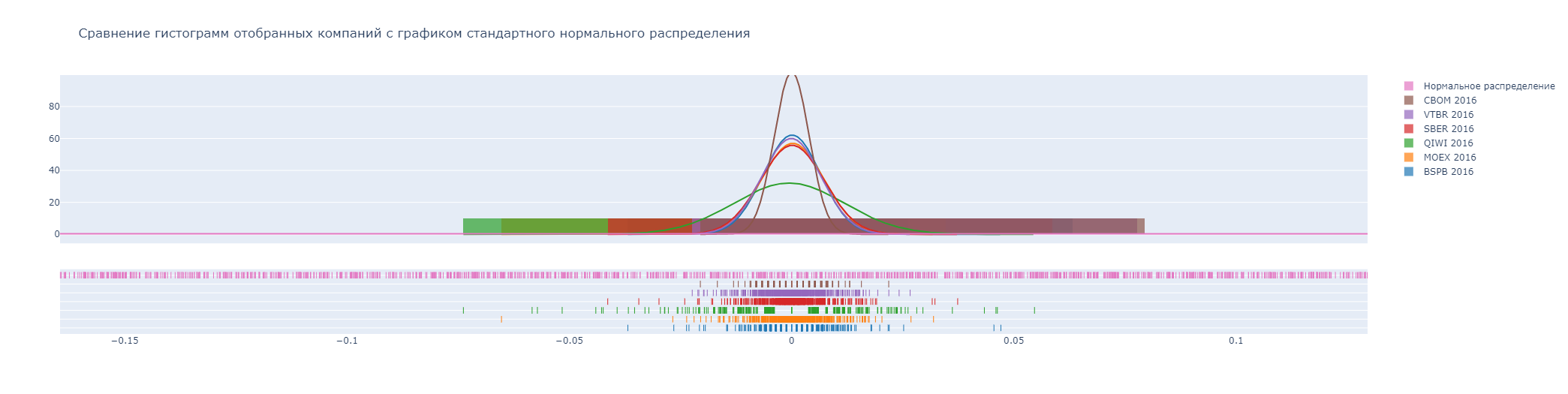


Рисунок 13 - Сравнение гистограмм отобранных компаний с графиком стандартного нормального распределения в случае полного года[28]

Из гистограммы следует, что выбранные периоды немного схожи с нормальным распределением, выделенным синим цветом, хотя имеют не свойственные ему значения. Логдоходность ПАО «Банк Санкт-Петербург»,сосредоточена в отрезке [-0.07;0.07], а в основном все компании находятся на этом промежутке также сильно вытянувшись, подробнее можно посмотреть в Jupyter Notebook пользуясь преимуществами интерактивных графиков библиотеки plotly.

**Заключение**

В процессе написания данной работы были выполнены все поставленные задачи, а также соблюдены все требования её выполнения. Было выяснено, что критерий Хегази-Грина достаточно удобен в использовании, а также имеет довольно высокую мощность, что позволяет получить наиболее точный результат. Исследование динамики курса акций компаний, которые входят в индекс МосБиржи финансы за период с 2014 по 2020 года. Проведённые расчёты показали, что гипотеза о нормальности распределения логарифмических доходностей была отвергнута на трех промежутках и таймфреймах, после проверки 4 разными способами, а именно с помощью первого и второго критерия Хегази-Грина, критерия Колмогорова, а также графического метода, при этом проверка на модельных данных дала положительный результат, что свидетельствует о правильности рассуждений. То есть реальных на данных мы пришли к отторжению гипотезы, то есть наблюдается отклонение логарифмических доходностей котировок акций компании, которые входят в индекс ММВБ финансы.

Можно сделать однозначный вывод, что при включении как параметр логарифмическую доходность в обучение нейронных сетей на биржевых данных, а точнее на котировках, которые были рассмотрены в данной курсовой работе, можно не проверять на нормальность, как в основном логарифмическая доходность не распределена нормально. Поэтому приведение столбца логарифмической доходности с помощью инструментов и возможностей языка программирования Python к нормальному распределению, без предварительной проверки на нормальность, является оптимальным. Тем самым мы можем сэкономить время аналитиков данных, занимающих данными котировками и настраивающие нейронную сеть.

Важно отметить, что вычисления и их алгоритмы правильны, так как гипотеза о нормальном распределении логарифмических доходностей подтверждается на модельных данных. Также о правильности произведенных вычислений может говорить тот факт, что анализ данных был произведен для большого количества проверок и статистик.

Также новизну работы можно оценить в использовании интерактивных графиков библиотеки plotly. Которые позволяют быстро оценить распределение по каждой котировке и году. Тем самым

В заключение, хотелось бы сделать акцент на том, что нельзя представить использование математики для анализа финансовых рынков без связи с реальными данными и процессами. И в настоящей работе моно увидеть то, как применяется аппарат математической статистики к процессам образования цен.

**СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ**

**ЛИТЕРАТУРА**

1. Браилов А.В. Лекции по математической статистике. – М.: Финакадемия, 2007.

2. Красс М. С., Б.П. Чупрынов Б.П. Математика в экономике - М.: Финансы и статистика,2007.

3. Колемаев В.А., Калинина В.Н. Теория вероятностей и математическая статистика. М.: КноРус, 2017.

4. Кацко И.А. Теория вероятностей и математическая статистика (для бакалавров). Учебное пособие. М.: КноРус, 2019.

5. Глебов Криволапов Практикум по математической статистике. Проверка гипотез с использованием Excel, MatCale, R и Python. М.: Прометей, 2019.

6. Малугин В.А. Математическая статистика. М.: Юрайт, 2020.

7. Карлов А. М. Теория вероятностей и математическая статистика для экономистов. М.: КноРус, 2020.

8. Энатская Н. Ю Математическая статистика и случайные процессы. М.: Юрайт, 2020.

**Интернет ресурсы**

1. <https://www.finam.ru/>  (дата обращения: 26.04.2021).
2. Сайт Московской биржи <https://www.moex.com/>

Приложения

Приложение 1 Модельные данные

Прежде, чем приступать к основному анализу данной курсовой работы необходимо удостовериться в правильности работы выбранного нами алгоритма на модельных данных, распределенных по нормальному закону (объем выборки равен 250, что приблизительно равно количеству торговых дней за год). Для этого будем использовать программу, в которой методом Монте-Карло (5000000 испытаний) будут рассчитаны 999 квантилей распределения статистики критерия Гири для модельных данных. В таблице 17 представлены только 9 квантилей (0.1, 0.2, … , 0.9), остальные данные можно найти в файле «1000quantiles(n=250)»:

|  |  |
| --- | --- |
| **q** | **x\_q** |
| 0,1 | 1,075718 |
| 0,2 | 1,093935 |
| 0,3 | 1,107228 |
| 0,4 | 1,118368 |
| 0,5 | 1,128688 |
| 0,6 | 1,138248 |
| 0,7 | 1,149202 |
| 0,8 | 1,162234 |
| 0,9 | 1,17964 |

|  |  |
| --- | --- |
| **q** | **x\_q** |
| 0,1 | 1,115717 |
| 0,2 | 1,123938 |
| 0,3 | 1,137228 |
| 0,4 | 1,148368 |
| 0,5 | 1,158689 |
| 0,6 | 1,168248 |
| 0,7 | 1,173206 |
| 0,8 | 1,1767 |
| 0,9 | 1,186 |

Табл. 17.Таблицы квантилей для статистик Т1 и Т2

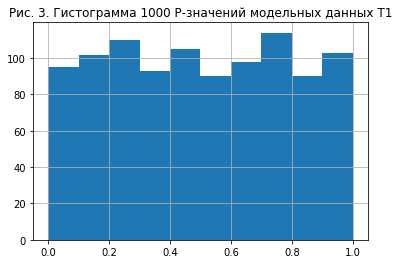
Теперь для сгенерированных модельных данных необходимо вычислить 1000 p-значений и произвести проверку равномерности распределения этих значений на отрезке [0;1] используя вспомогательный критерий Колмогорова. Для наглядности производится построение гистограммы.

Рисунок 23. Гистограмма 1000 p-значений модельных данных для статистики T1(на нормально распределенной выборке).

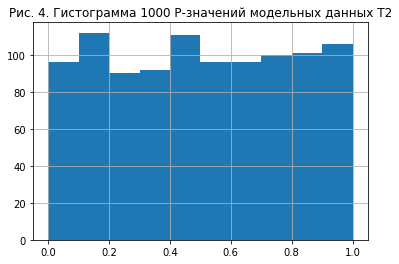


Рисунок 24. Гистограмма 1000 p-значений модельных данных для статистики T2(на нормально распределенной выборке).

Гистограммы наглядно демонстрируют, что p-значения на заданном отрезке распределены равномерно. Это подтверждает и вычисленное p-значение критерия Колмогорова-Смирнова, равное 0.95326 для Т1 и 0.97804 для Т2, что больше уровня значимости 0.05. Следовательно, мы принимаем гипотезу о нормальном распределении и можем перейти к проверке критерия на реальных данных.

Приложение 2

Технические характеристики компьютера:

Процессор Intel(R) Core(tm) i5-8250U CPU @

Тактовая частота 1.60 GHz 1.80 GHz

Частота системной шины 4 GT/s OPI

Объём кэша второго уровня 1,0 Мб

Приложение 2.

Код программ:

In [1]# Подключение библиотек

import time

import pandas as pd

import datetime as dt

import matplotlib.pyplot as plt

import scipy.stats

import numpy as np

from pynverse import inversefunc

from scipy.stats import norm

import plotly.express as px

import plotly.graph\_objects as go

import plotly.figure\_factory as ff

from plotly.offline import iplot

from scipy.stats import mstats, laplace

import math

from scipy.stats import \*

In [2]

# Предоброботка данных для 2014-2018 день

%cd /content/Probability-theory/Курсач/2014-2018 день

In [3]

start\_time = time.time()

tickers = ['BSPB','CBOM','MOEX','QIWI','SBER','SFIN','VTBR']

data = []

for i in tickers:

  data.append(pd.read\_csv(i+'\_day\_2014-2018.csv'))

for i in data:

  del i['<TIME>']

  i['<DATE>'] = pd.to\_datetime(i['<DATE>'], format = '%Y%m%d')

  i['<JUMP>'] = (i['<CLOSE>'] - i['<CLOSE>'].shift(1))/i['<CLOSE>']

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд') # Время выполнения: 0.04 секунд

In [4]

start\_time = time.time()

frames = []

for i in data:

  frames.append(i.groupby(i['<DATE>'].dt.year)['<DATE>'].agg(['count']))

result = pd.concat(frames, axis=1)

result.columns = tickers

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд') # Время выполнения: 0.03 секунд

In [5]#

result

start\_time = time.time()

tickers = ['BSPB','MOEX','QIWI','SBER','VTBR']

data = []

for i in tickers:

  data.append(pd.read\_csv(i+'\_day\_2014-2018.csv'))

for i in data:

  del i['<TIME>']

  i['<DATE>'] = pd.to\_datetime(i['<DATE>'], format = '%Y%m%d')

  i['<JUMP>'] = (i['<CLOSE>'] - i['<CLOSE>'].shift(1))/i['<CLOSE>']

frames = []

for i in data:

  frames.append(i.groupby(i['<DATE>'].dt.year)['<DATE>'].agg(['count']))

result = pd.concat(frames, axis=1)

result.columns = tickers

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд') # Время выполнения: 0.03 секунд

In [6]#

start\_time = time.time()

tickers = ['BSPB','MOEX','QIWI','SBER','VTBR']

data = []

for i in tickers:

  data.append(pd.read\_csv(i+'\_day\_2014-2018.csv'))

for i in data:

  del i['<TIME>']

  i['<DATE>'] = pd.to\_datetime(i['<DATE>'], format = '%Y%m%d')

  i['<JUMP>'] = (i['<CLOSE>'] - i['<CLOSE>'].shift(1))/i['<CLOSE>']

frames = []

for i in data:

  frames.append(i.groupby(i['<DATE>'].dt.year)['<DATE>'].agg(['count']))

result = pd.concat(frames, axis=1)

result.columns = tickers

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд') # Время выполнения: 0.03 секунд

start\_time = time.time()

uframes = []

dframes = []

for i in data:

  uframes.append(i.groupby(i['<DATE>'].dt.year)['<JUMP>'].agg(['max']))

  dframes.append(i.groupby(i['<DATE>'].dt.year)['<JUMP>'].agg(['min']))

uresult = pd.concat(uframes, axis=1)

uresult.columns = tickers

dresult = pd.concat(dframes, axis=1)

dresult.columns = tickers

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')  # Время выполнения: 0.06 секунд

start\_time = time.time()

plt.rcParams["figure.figsize"] = (10,6)

plt.rcParams['figure.dpi'] = 120

ticker = data[0] # BSPB

points = pd.to\_numeric(ticker.groupby(ticker['<DATE>'].dt.year)['<JUMP>'].agg(['idxmax'])['idxmax'])

plt.plot(ticker['<DATE>'], ticker['<CLOSE>'], label='BSPB')

for i in points:

  plt.plot([ticker['<DATE>'][i-1],ticker['<DATE>'][i]], [ticker['<CLOSE>'][i-1], ticker['<CLOSE>'][i]], '^:', color='lime', linewidth=3, markersize=6)

plt.title("Максимальные дневные относительные скачки цен вверх Банк Санкт-Петербурга")

plt.xticks(rotation=70)

plt.legend()

plt.show()

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд') # Время выполнения: 0.46 секунд

start\_time = time.time()

plt.rcParams["figure.figsize"] = (10,6)

plt.rcParams['figure.dpi'] = 120

ticker = data[0] # BSPB

points = pd.to\_numeric(ticker.groupby(ticker['<DATE>'].dt.year)['<JUMP>'].agg(['idxmin'])['idxmin'])

plt.plot(ticker['<DATE>'], ticker['<CLOSE>'], label='BSPB')

for i in points:

  plt.plot([ticker['<DATE>'][i-1],ticker['<DATE>'][i]], [ticker['<CLOSE>'][i-1], ticker['<CLOSE>'][i]], 'v:', color='red', linewidth=3, markersize=6)

plt.title("Максимальные дневные относительные скачки цен вниз Банк Санкт-Петербурга")

plt.xticks(rotation=70)

plt.legend()

plt.show()

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд') # Время выполнения: 0.32 секунд

#Проверка гипотез для модельных данных

In [10]#

##Квантили и P-значения T1

start\_time = time.time()

m = 5000000

n = 250

T1\_list = []

def mu(i,n):

 return(4.91\*(((i)/(n+1))\*\*0.14-((n-i+1)/(n+1))\*\*0.14))

def T1\_stat(n):

 viborka = np.random.normal(loc = 0, scale = 1, size = n)

 summ = 0

 x\_mean = np.mean(viborka)

 s2 = np.var(viborka)\*n/(n-1)

 for i in range(n):

  z = (viborka[i]-x\_mean)/s2\*\*(1/2)

  nu = mu(i,n)

  summ+=abs(z-nu)

 return summ/n

for i in range(m): #Метод Монте-карло для поиска стандартной ошибки и значений статистики

 T1 = T1\_stat(250)

 T1\_list.append(T1)

quant = [el for el in np.arange(0.1,1,0.1)]

quant1 = [el for el in np.arange(0.001,1,0.001)]

ll = []

for i in quant: # вычисление квантилей

 ll.append(np.quantile(T1\_list,i))

df9Quant1 = pd.DataFrame({'Квантиль':quant, 'Значение':ll}) # Создание таблицы 9 квантилей

ll1 = []

for i in quant1: # вычисление квантилей

 ll1.append(np.quantile(T1\_list,i))

df999Quant1 = pd.DataFrame({'Квантиль':quant1, 'Значение':ll1}) # Создание таблицы 999 квантилей

df999Quant1.to\_csv('1000quantilesT1.csv')

def p\_val1(k): # Создание функции для вычисления р-значения

 w = T1\_stat(k) # Вычисление статистики для данной выборки

 s = 0

 for i in range(len(ll1)):

  if w > ll1[i]:

    s+=1

 return round(s/len(quant1),3)

lll = []

for i in range(1000):

 y = p\_val1(250)

 lll.append(y)

level = [0.1\*i for i in range(1, 10)]

q = [np.quantile(T1\_list , i) for i in level]

dict\_q = {"q":level,"x\_q":q}

Z\_q = pd.DataFrame(dict\_q)

## сохранение квантилей в виде таблицы Excel

Z\_q.to\_excel('10 Квантилей объем выборки 250 для T1.xls')

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

##Гистограмма P-значений МД Т1

start\_time = time.time()

plt.hist(lll) # построение гистограммы Р-значений на отрезке от 0 до 1

plt.grid(True)

plt.title('Рис. 3. Гистограмма 1000 Р-значений модельных данных Т1')

plt.show()

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

##Проверка критерием Колмогорова

start\_time = time.time()

pvalue = kstest(lll,'uniform')

print(round(pvalue[1],5))

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

#Квантили и P-значения T2

start\_time = time.time()

m = 5000000

n = 250

T2\_list = []

def mu(i,n):

 return(4.91\*(((i)/(n+1))\*\*0.14-((n-i+1)/(n+1))\*\*0.14))

def T2\_stat(n):

 viborka = np.random.normal(loc = 0, scale = 1, size = n)

 summ = 0

 x\_mean = np.mean(viborka)

 s2 = np.var(viborka)\*n/(n-1)

 for i in range(n):

  z = (viborka[i]-x\_mean)/s2\*\*(1/2)

  nu = mu(i,n)

  summ+=(z-nu)\*\*2

 return summ/n

for i in range(m): #Метод Монте-карло для поиска стандартной ошибки и значений статистики

 T2 = T2\_stat(250)

 T2\_list.append(T2)

quant = [el for el in np.arange(0.1,1,0.1)]

quant1 = [el for el in np.arange(0.001,1,0.001)]

ll = []

for i in quant: # вычисление квантилей

 ll.append(np.quantile(T2\_list,i))

df9Quant1 = pd.DataFrame({'Квантиль':quant, 'Значение':ll}) #Создание таблицы 9 квантилей

ll1 = []

for i in quant1: # вычисление квантилей

 ll1.append(np.quantile(T2\_list,i))

df999Quant1 = pd.DataFrame({'Квантиль':quant1, 'Значение':ll1})# Создание таблицы 999 квантилей

df999Quant1.to\_csv('1000quantilesT2.csv')

def p\_val2(k): # Создание функции для вычисления р-значения

 w = T2\_stat(k) # Вычисление статистики для данной выборки

 s = 0

 for i in range(len(ll1)):

  if w > ll1[i]:

    s+=1

 return round(s/len(quant1),3)

lll = []

for i in range(1000):

 y = p\_val2(250)

 lll.append(y)

level = [0.1\*i for i in range(1, 10)]

q = [np.quantile(T1\_list , i) for i in level]

dict\_q = {"q":level,"x\_q":q}

Z\_q = pd.DataFrame(dict\_q)

## сохранение квантилей в виде таблицы Excel

Z\_q.to\_excel('10 Квантилей объем выборки 250 для T2.xls')

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

##Гистограмма P-значений МД Т2

start\_time = time.time()

plt.hist(lll) # построение гистограммы Р-значений на отрезке от 0 до 1

plt.grid(True)

plt.title('Рис. 4. Гистограмма 1000 Р-значений модельных данных Т2')

plt.show()

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

##Проверка критерием Колмогорова

start\_time = time.time()

pvalue = kstest(lll,'uniform')

print(round(pvalue[1],5))

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

#Мощность критерия

##Для первой статистики[14]

import scipy.stats as st

import numpy as np

import pandas as pd

n = 261

p = [0.01, 0.05, 0.15]

qs = pd.read\_csv('1000quantilesT1.csv') # загрузка файла 1000квантилей

qs = qs['Значение']

qs = list(qs)

A = 100

def elaw(x):

 return[x.count(i) for i in set(x)]

def m(R):

  return [norm.ppf(i / (len(R) + 1)) for i in range(1, len(R) + 1)]

def pl(x, a):

 return len([i for i in x if i < a]) / len(x)

def pg(x, a):

 return len([i for i in x if i > a]) / len(x)

df = pd.DataFrame(index = [str(int(i \* 100)) + '%' for i in p])

pv1 = []

pr1 = []

pv2 = []

pr2 = []

for i in range(A):

 R = np.random.standard\_t(8, size = n)

 aR = sum(R) / n

 s2 = sum((R - aR) \*\* 2) / (n - 1)

 Z = (R - aR) / (s2 \*\* 0.5)

 M = m(R)

 T1 = sum(abs(Z - M)) / n

 T2 = sum((Z - M) \*\* 2) / n

 pv1.append(pl(qs, T1))

 pv2.append(pl(qs, T2))

pr1.append(pg(pv1, p[0]))

pr1.append(pg(pv1, p[1]))

pr1.append(pg(pv1, p[2]))

df['T1'] = pr1

df = df.T

df.to\_excel('power\_T1.xlsx')

##Для второй статистики

import scipy.stats as st

import numpy as np

import pandas as pd

n = 261

p = [0.01, 0.05, 0.15]

qs = pd.read\_csv('1000quantilesT2.csv') # загрузка файла 1000квантилей

qs = qs['Значение']

qs = list(qs)

A = 100

def elaw(x):

 return[x.count(i) for i in set(x)]

def m(R):

  return [norm.ppf(i / (len(R) + 1)) for i in range(1, len(R) + 1)]

def pl(x, a):

 return len([i for i in x if i < a]) / len(x)

def pg(x, a):

 return len([i for i in x if i > a]) / len(x)

df = pd.DataFrame(index = [str(int(i \* 100)) + '%' for i in p])

pv2 = []

pr2 = []

for i in range(A):

 R = np.random.standard\_t(8, size = n)

 aR = sum(R) / n

 s2 = sum((R - aR) \*\* 2) / (n - 1)

 Z = (R - aR) / (s2 \*\* 0.5)

 M = m(R)

 T2 = sum((Z - M) \*\* 2) / n

 pv2.append(pl(qs, T2))

pr2.append(pg(pv2, p[0]))

pr2.append(pg(pv2, p[1]))

pr2.append(pg(pv2, p[2]))

df['T2'] = pr2

df = df.T

df.to\_excel('power\_T2.xlsx')

##Проверка гипотезы для реальных данных

In [20]

start\_time = time.time()

qs = pd.read\_csv('1000quantilesT1.csv')# загрузка файла 1000 квантилей

qs = qs['Значение']

qs = list(qs)

F = (lambda y: norm.cdf(x = y,loc = 0,scale = 1))

invF = inversefunc(F)

def T1\_stat(year,file):

  df = pd.read\_csv(file)

  condition = (df['<DATE>']>= year\*10000) & (df['<DATE>']<(year+1)\*10000) #условие на исследуемый период

  df = df[condition]['<CLOSE>']

  pd.to\_numeric(df)

  df = 100 \* np.log(df).diff()[1:] # расчет логдоходностей

  df = list(df)

  n = len(df)

  summ = 0

  x\_mean = np.mean(df)

  s2 = np.var(df)\*n/(n-1)

  for i in range(n):

    z = (df[i]-x\_mean)/s2\*\*(1/2)

    nu = invF((i+1)/(n+1))

    summ+=abs(z-nu)

  return summ/n

def p\_val1(year,file): # Создание функции для вычисления р-значения

  w = T1\_stat(year,file) # Вычисление статистики для данной выборки

  s = 0

  for i in range(len(qs)):

    if w > qs[i]:

      s+=1

  return round(s/len(qs),3)

tickers =['BSPB','CBOM','MOEX','QIWI','SBER','SFIN','VTBR']

years = range(2018,2021)

df\_P\_value = pd.DataFrame()

df\_P\_value['Тикер'] = tickers

p\_val\_list = []

for year in years: #заполнение даты фрейм Р-значениями

  yearP\_val = []

  for ticker in tickers:

    yearP\_val.append((p\_val1(year, ticker + '\_day\_2018-2020.csv')))

    p\_val\_list.append(p\_val1(year, ticker + '\_day\_2018-2020.csv'))

  df\_P\_value[str(year)] = yearP\_val

df\_P\_value.to\_excel('real\_T1\_2018-2020.xlsx')

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

In [17]

start\_time = time.time()

ll1 = []

counter=0

pictures = 5

tickers = ['BSPB','MOEX','QIWI','SBER','VTBR']

for i in tickers:

    ll1 = list(df\_P\_value.iloc[counter])

    plt.hist(ll1[1:])

    plt.grid(True)

    pictures\_1 = str(pictures)

    company =str(ll1[0])

    title = 'Рис.'+pictures\_1+ '. '+'Гистограмма Р-значений за 2014-2018 (T1): '+ company

    pvalue = kstest(ll1[1:],'uniform')[1]

    print(pvalue)

    plt.title(title)

    ll1=[]

    counter+=1

    pictures+= 1

    plt.show()

#Критерий Колмогорова

pvalue = ks\_2samp(p\_val\_list,qs)

pvalue

#Проверка на реальных данных Т2

In [21]

start\_time = time.time()

qs = pd.read\_csv('1000quantilesT2.csv')# загрузка файла 1000 квантилей

qs = qs['Значение']

qs = list(qs)

F = (lambda y: norm.cdf(x = y,loc = 0,scale = 1))

invF = inversefunc(F)

def T1\_stat(year,file):

  df = pd.read\_csv(file)

  condition = (df['<DATE>']>= year\*10000) & (df['<DATE>']<(year+1)\*10000) #условие на исследуемый период

  df = df[condition]['<CLOSE>']

  pd.to\_numeric(df)

  df = 100 \* np.log(df).diff()[1:] # расчет логдоходностей

  df = list(df)

  n = len(df)

  summ = 0

  x\_mean = np.mean(df)

  s2 = np.var(df)\*n/(n-1)

  for i in range(n):

    z = (df[i]-x\_mean)/s2\*\*(1/2)

    nu = invF((i+1)/(n+1))

    summ+=abs(z-nu)\*\*2

  return summ/n

def p\_val1(year,file): # Создание функции для вычисления р-значения

  w = T1\_stat(year,file) # Вычисление статистики для данной выборки

  s = 0

  for i in range(len(qs)):

    if w > qs[i]:

      s+=1

  return round(s/len(qs),3)

tickers =['BSPB','CBOM','MOEX','QIWI','SBER','SFIN','VTBR']

years = range(2018,2021)

df\_P\_value = pd.DataFrame()

df\_P\_value['Тикер'] = tickers

p\_val\_list = []

for year in years: #заполнение даты фрейм Р-значениями

  yearP\_val = []

  for ticker in tickers:

    yearP\_val.append((p\_val1(year, ticker + '\_day\_2018-2020.csv')))

    p\_val\_list.append(p\_val1(year, ticker + '\_day\_2018-2020.csv'))

  df\_P\_value[str(year)] = yearP\_val

df\_P\_value.to\_excel('real\_T2\_2018-2020.xlsx')

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

In [31]

start\_time = time.time()

ll1 = []

years = range(2018,2021)

tickers =['BSPB','CBOM','MOEX','QIWI','SBER','SFIN','VTBR']

for year in years:

 for i in range(5):

  ll1.append(df\_P\_value[str(year)][i])

plt.hist(ll1)

plt.grid(True)

plt.title('Рис. 7. Гистограмма Р-значений за 2018-2020 (T1)')

plt.show()

## Графический метод

In [23]

BSPB = pd.read\_csv("BSPB\_day\_2014-2018.csv")

MOEX = pd.read\_csv("MOEX\_day\_2014-2018.csv")

QIWI = pd.read\_csv("QIWI\_day\_2014-2018.csv")

SBER = pd.read\_csv("SBER\_day\_2014-2018.csv")

VTBR= pd.read\_csv("VTBR\_day\_2014-2018.csv")

BSPB['YEAR']  = (round(BSPB['<DATE>']/10000,0))

MOEX['YEAR']  = (round(MOEX ['<DATE>']/10000,0))

QIWI['YEAR']  = (round(QIWI['<DATE>']/10000,0))

SBER['YEAR']  = (round(SBER['<DATE>']/10000,0))

VTBR['YEAR']  = (round(VTBR['<DATE>']/10000,0))

x1 = list(np.log((BSPB.loc[BSPB['YEAR'] == 2014]["<CLOSE>"]).pct\_change()+1))

x2 = list(np.log((MOEX.loc[MOEX['YEAR'] == 2014]["<CLOSE>"]).pct\_change()+1))

x3 = list(np.log((QIWI.loc[QIWI['YEAR'] == 2014]["<CLOSE>"]).pct\_change()+1))

x4 = list(np.log((SBER.loc[SBER['YEAR'] == 2014]["<CLOSE>"]).pct\_change()+1))

x5 = list(np.log((VTBR.loc[VTBR['YEAR'] == 2014]["<CLOSE>"]).pct\_change()+1))

x6 = list(np.log((BSPB.loc[BSPB['YEAR'] == 2015]["<CLOSE>"]).pct\_change()+1))

x7 = list(np.log((MOEX.loc[MOEX['YEAR'] == 2015]["<CLOSE>"]).pct\_change()+1))

x8 = list(np.log((QIWI.loc[QIWI['YEAR'] == 2015]["<CLOSE>"]).pct\_change()+1))

x9 = list(np.log((SBER.loc[SBER['YEAR'] == 2015]["<CLOSE>"]).pct\_change()+1))

x10 = list(np.log((VTBR.loc[VTBR['YEAR'] == 2015]["<CLOSE>"]).pct\_change()+1))

x11 = list(np.log((BSPB.loc[BSPB['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x12= list(np.log((MOEX.loc[MOEX['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x13= list(np.log((QIWI.loc[QIWI['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x14 = list(np.log((SBER.loc[SBER['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x15 = list(np.log((VTBR.loc[VTBR['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x16 = list(np.log((BSPB.loc[BSPB['YEAR'] == 2017]["<CLOSE>"]).pct\_change()+1))

x17 = list(np.log((MOEX.loc[MOEX['YEAR'] == 2017]["<CLOSE>"]).pct\_change()+1))

x18= list(np.log((QIWI.loc[QIWI['YEAR'] == 2017]["<CLOSE>"]).pct\_change()+1))

x19= list(np.log((SBER.loc[SBER['YEAR'] == 2017]["<CLOSE>"]).pct\_change()+1))

x20= list(np.log((VTBR.loc[VTBR['YEAR'] == 2017]["<CLOSE>"]).pct\_change()+1))

del x1[0]

del x2[0]

del x3[0]

del x4[0]

del x5[0]

del x6[0]

del x7[0]

del x8[0]

del x9[0]

del x10[0]

del x11[0]

del x12[0]

del x13[0]

del x14[0]

del x15[0]

del x16[0]

del x17[0]

del x18[0]

del x19[0]

del x20[0]

norm = np.random.normal(loc=0.0, scale=1.0, size=10000)

hist\_data = [x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13,x14,x15,x16,x17,x18,x19,x20,norm]

group\_labels = ["BSPB 2014","MOEX 2014","QIWI 2014","SBER 2014","VTBR 2014","BSPB 2015","MOEX 2015","QIWI 2015","SBER 2015","VTBR 2015","BSPB 2016","MOEX 2016","QIWI 2016","SBER 2016","VTBR 2016","BSPB 2017","MOEX 2017","QIWI 2017","SBER 2017","VTBR 2017", 'Нормальное распределение']

fig = ff.create\_distplot(hist\_data, group\_labels,curve\_type = 'normal',bin\_size=.1)

fig.update\_layout(title\_text='Cравнение гистограмм отобранных компаний с графиком стандартного нормального распределения')

fig.show()

#Проверка на реальных данных Т1 для часа

In [24]

start\_time = time.time()

qs = pd.read\_csv('1000quantilesT1.csv')# загрузка файла 1000 квантилей

qs = qs['Значение']

qs = list(qs)

F = (lambda y: norm.cdf(x = y,loc = 0,scale = 1))

invF = inversefunc(F)

def T1\_stat(year,file):

  df = pd.read\_csv(file)

  condition = (df['<DATE>']>= year\*10000) & (df['<DATE>']<(year+1)\*10000) #условие на исследуемый период

  df = df[condition]['<CLOSE>']

  pd.to\_numeric(df)

  df = 100 \* np.log(df).diff()[1:] # расчет логдоходностей

  df = list(df)

  n = len(df)

  summ = 0

  x\_mean = np.mean(df)

  s2 = np.var(df)\*n/(n-1)

  for i in range(n):

    z = (df[i]-x\_mean)/s2\*\*(1/2)

    nu = invF((i+1)/(n+1))

    summ+=abs(z-nu)

  return summ/n

def p\_val1(year,file): # Создание функции для вычисления р-значения

  w = T1\_stat(year,file) # Вычисление статистики для данной выборки

  s = 0

  for i in range(len(qs)):

    if w > qs[i]:

      s+=1

  return round(s/len(qs),3)

tickers =['BSPB','CBOM','MOEX','QIWI','SBER','VTBR']

year= 2016

df\_P\_value = pd.DataFrame()

df\_P\_value['Тикер'] = tickers

p\_val\_list = []

for ticker in tickers:

  p\_val\_list.append(p\_val1(year, ticker + '\_hour\_2014-2018.csv'))

df\_P\_value["2016"] = p\_val\_list

df\_P\_value.to\_excel('real\_T1\_2016-2016.xlsx')

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

In [25]

start\_time = time.time()

ll1 = []

tickers =['BSPB','CBOM','MOEX','QIWI','SBER','VTBR']

for i in range(5):

  ll1.append(df\_P\_value["2016"][i])

plt.hist(ll1)

plt.grid(True)

plt.title('Рис. 7. Гистограмма Р-значений за 2016-2016 (T1)')

plt.show()

In [26]

start\_time = time.time()

qs = pd.read\_csv('1000quantilesT2.csv')# загрузка файла 1000 квантилей

qs = qs['Значение']

qs = list(qs)

F = (lambda y: norm.cdf(x = y,loc = 0,scale = 1))

invF = inversefunc(F)

def T1\_stat(year,file):

  df = pd.read\_csv(file)

  condition = (df['<DATE>']>= year\*10000) & (df['<DATE>']<(year+1)\*10000) #условие на исследуемый период

  df = df[condition]['<CLOSE>']

  pd.to\_numeric(df)

  df = 100 \* np.log(df).diff()[1:] # расчет логдоходностей

  df = list(df)

  n = len(df)

  summ = 0

  x\_mean = np.mean(df)

  s2 = np.var(df)\*n/(n-1)

  for i in range(n):

    z = (df[i]-x\_mean)/s2\*\*(1/2)

    nu = invF((i+1)/(n+1))

    summ+=abs(z-nu)\*\*2

  return summ/n

def p\_val1(year,file): # Создание функции для вычисления р-значения

  w = T1\_stat(year,file) # Вычисление статистики для данной выборки

  s = 0

  for i in range(len(qs)):

    if w > qs[i]:

      s+=1

  return round(s/len(qs),3)

tickers =['BSPB','CBOM','MOEX','QIWI','SBER','VTBR']

year= 2016

df\_P\_value = pd.DataFrame()

df\_P\_value['Тикер'] = tickers

p\_val\_list = []

for ticker in tickers:

  p\_val\_list.append(p\_val1(year, ticker + '\_hour\_2014-2018.csv'))

df\_P\_value["2016"] = p\_val\_list

df\_P\_value.to\_excel('real\_T2\_2016-2016.xlsx')

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

In [27]

start\_time = time.time()

ll1 = []

counter=0

pictures = 10

tickers = ['BSPB','MOEX','QIWI','SBER','VTBR']

for i in tickers:

    ll1 = list(df\_P\_value.iloc[counter])

    plt.hist(ll1[1:])

    plt.grid(True)

    pictures\_1 = str(pictures)

    company =str(ll1[0])

    title = 'Рис.'+pictures\_1+ '. '+'Гистограмма Р-значений за 2014-2018 (T2): '+ company

    pvalue = kstest(ll1[1:],'uniform')[1]

    print(pvalue)

    plt.title(title)

    ll1=[]

    counter+=1

    pictures+= 1

    plt.show()

In [28]

BSPB = pd.read\_csv("BSPB\_hour\_2014-2018.csv")

CBOM = pd.read\_csv("CBOM\_hour\_2014-2018.csv")

MOEX = pd.read\_csv("MOEX\_hour\_2014-2018.csv")

QIWI = pd.read\_csv("QIWI\_hour\_2014-2018.csv")

SBER = pd.read\_csv("SBER\_hour\_2014-2018.csv")

VTBR= pd.read\_csv("VTBR\_hour\_2014-2018.csv")

BSPB['YEAR']  = (round(BSPB['<DATE>']/10000,0))

MOEX['YEAR']  = (round(MOEX ['<DATE>']/10000,0))

QIWI['YEAR']  = (round(QIWI['<DATE>']/10000,0))

SBER['YEAR']  = (round(SBER['<DATE>']/10000,0))

VTBR['YEAR']  = (round(VTBR['<DATE>']/10000,0))

CBOM['YEAR']  = (round(BSPB['<DATE>']/10000,0))

x1 = list(np.log((BSPB.loc[BSPB['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x2 = list(np.log((MOEX.loc[MOEX['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x3 = list(np.log((QIWI.loc[QIWI['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x4 = list(np.log((SBER.loc[SBER['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x5 = list(np.log((VTBR.loc[VTBR['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

x6 = list(np.log((CBOM.loc[VTBR['YEAR'] == 2016]["<CLOSE>"]).pct\_change()+1))

del x1[0]

del x2[0]

del x3[0]

del x4[0]

del x5[0]

del x6[0]

norm = np.random.normal(loc=0.0, scale=1.0, size=10000)

hist\_data = [x1, x2, x3, x4, x5, x6,norm]

group\_labels = ["BSPB 2016","MOEX 2016","QIWI 2016","SBER 2016","VTBR 2016","CBOM 2016",'Нормальное распределение']

fig = ff.create\_distplot(hist\_data, group\_labels,curve\_type = 'normal',bin\_size=.1)

fig.update\_layout(title\_text='Cравнение гистограмм отобранных компаний с графиком стандартного нормального распределения')

fig.show()

In [29]

start\_time = time.time()

def p\_val1(year,file):

  df = pd.read\_csv(file)

  condition = (df['<DATE>']>= year\*10000) & (df['<DATE>']<(year+1)\*10000) #условие на исследуемый период

  df = df[condition]['<CLOSE>']

  pd.to\_numeric(df)

  df = 100 \* np.log(df).diff()[1:] # расчет логдоходностей

  df = list(df)

  return kstest(df, 'norm')[1]

tickers = ['BSPB','CBOM','MOEX','QIWI','SBER','SFIN','VTBR']

years = range(2018,2021)

df\_P\_value = pd.DataFrame()

df\_P\_value['Тикер'] = tickers

p\_val\_list = []

for year in years: #заполнение даты фрейм Р-значениями

  yearP\_val = []

  for ticker in tickers:

    yearP\_val.append((p\_val1(year, ticker + '\_day\_2018-2020.csv')))

    p\_val\_list.append(p\_val1(year, ticker + '\_day\_2018-2020.csv'))

  df\_P\_value[str(year)] = yearP\_val

df\_P\_value.to\_excel('real\_Kolomogorov\_2018-2020.xlsx')

print('Время выполнения:', round(time.time()-start\_time, 2), 'секунд')

Приложение №3

Курсовая\_быханов.ipynb

1000quantilesT1.csv

1000quantilesT2.csv

10 Квантилей объем выборки 250 для T1.xlsx

10 Квантилей объем выборки 250 для T2.xlsx

real\_Kolomogorov\_2018-2020.xlsx

SBER\_day\_2014-2018.csv

BSPB\_day\_2014-2018.csv

SFIN\_day\_2014-2018.csv

CBOM\_day\_2014-2018.csv

VTBR\_day\_2014-2018.csv

MOEX\_day\_2014-2018.csv

QIWI\_day\_2014-2018.csv

MOEX\_day\_2018-2020.csv

SBER\_day\_2018-2020.csv

QIWI\_day\_2018-2020.csv

SFIN\_day\_2018-2020.csv

BSPB\_day\_2018-2020.csv

VTBR\_day\_2018-2020.csv

CBOM\_day\_2018-2020.csv

MOEX\_hour\_2014-2018.csv

SBER\_hour\_2014-2018.csv

QIWI\_hour\_2014-2018.csv

SFIN\_hour\_2014-2018.csv

BSPB\_hour\_2014-2018.csv

VTBR\_hour\_2014-2018.csv

CBOM\_hour\_2014-2018.csv

power.xlsx