Федеральное государственное образовательное бюджетное учреждение  
 высшего профессионального образования  
 «Финансовый университет при Правительстве Российской Федерации»

Департамент анализа данных, принятия решений и финансовых технологий

Курсовая работа

на тему

**«Проверка гипотезы** **о равенстве дисперсий логарифмической доходности индекса фондового рынка и входящих в его состав акций»**

Вид данных для исследования:

«**Котировки акций компаний, входящих в индекс ММВБ потребительского сектора**»

Выполнила:

студентка группы ПМ19-1,

Башмакова А. А.

Научный руководитель:

доцент, к.э.н.

Гринева Н. В.

Москва

2021

**Содержание**

[ВВЕДЕНИЕ 3](#_Toc72089922)

[I. ТЕОРЕТИЧЕСКАЯ СПРАВКА 5](#_Toc72089923)

[1. Математическая статистика 5](#_Toc72089924)

[2. Статистическая гипотеза 5](#_Toc72089925)

[3. Ошибки первого и второго рода 6](#_Toc72089926)

[4. P-значения 7](#_Toc72089927)

[5. Логарифмическая доходность 8](#_Toc72089928)

[6. Критерий Колмогорова 9](#_Toc72089929)

[7. Критерий Фишера 11](#_Toc72089930)

[II. ПРАКТИЧЕСКАЯ ЧАСТЬ 13](#_Toc72089931)

[1. Предварительный анализ данных 13](#_Toc72089932)

[2. Проверка гипотезы на модельных данных 17](#_Toc72089933)

[3. Проверка гипотезы на реальных данных 18](#_Toc72089934)

[4. Альтернативные гипотезы и оценка мощности критерия 21](#_Toc72089935)

[ЗАКЛЮЧЕНИЕ 22](#_Toc72089936)

[СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ 23](#_Toc72089937)

[ПРИЛОЖЕНИЯ 24](#_Toc72089938)

[Приложение 1 24](#_Toc72089939)

[Приложение 2 25](#_Toc72089940)

[Приложение 3 25](#_Toc72089941)

# **ВВЕДЕНИЕ**

В этой работе я проверю гипотезу о равенстве дисперсий логарифмической доходности фондового рынка и входящих в его состав акций. Для этого буду использовать критерий Фишера как относительно простой, но находящий широкое распространение в дисперсионном анализе.

Для исследования и проверки гипотезы я выбрала данные котировок акций тех компаний, которые входят в индекс ММВБ Московской биржи и относятся к потребительскому сектору – MOEXCN. Рассматривала я временной промежуток за последние 10 лет, то есть с 1 января 2011 года по 30 декабря 2021 года [9].

Индекс ММВБ представляет собой ценовый композитный фоновый индекс. Он считается одним из основных показателей экономики поскольку взвешен по рыночной капитализации. Он включает в себя акции крупнейших российских эмитентов, которые являются наиболее ликвидными и активно развивающимися и относятся к различным сферам жизни: сельское хозяйство, финансы, транспорт, промышленность, торговля, медицина.

Индекс MOEXCN (ММВБ потребительский сектор) на Московской бирже включает в себя компании, которые предоставляют различные продукты жизнедеятельности потребителям, например, обувь, продукты питания, игрушки, технику, алкоголь, а также ведут сельскохозяйственную деятельность. На апрель 2020 года в индекс входят 12 компаний: X5 Retail Group (компания, которая управляет продуктовыми магазинами «Пятёрочка», «Перекрёсток», «Чижик», «Карусель» ), ПАО «Детский мир» (сеть магазинов товаров для детей), ПАО «Магнит» (сеть магазинов под маркой «Магнит»: магазин «у дома», Супермаркет «Магнит Семейный», суперстор «Магнит Экстра», «Магнит Косметик», «Магнит Аптека»), ПАО «РУСАГРО» (самая крупная аграрная компания в России), ПАО "ЛЕНТА" (сеть продовольственный гипермаркетов первая по величине в России), МГП «Мать и Дитя» (сеть частных медицинских центров в области акушерства, гинекологии и педиатрии), ПАО «М.Видео» (сеть магазинов торговли бытовой техникой и электроникой), ПАО «Белуга Групп» (компания по производству алкогольных напитков), ПАО «Аптечная сеть 36,6» (сеть магазинов по продаже лекарственных средств), ПАО "Русская Аквакультура" (компания, которая занимается производством и продажей красной рыбы), ПАО «Со́ллерс» (компания, занимающаяся автомобилестроением), OR GROUP (торговая платформа, которая объединяет обувную сеть под брендом Westfalika.

Планируемая новизна заключается в проверке данной гипотезы с использованием языка программирования Python версии 3.0 и интерактивной оболочки для него Jupiter Notebook, благодаря которой возможно хранить и обрабатывать большой объём данных с использованием изображений, графиков, гистограмм, формул.

Актуальность работы заключается в использовании современного инструмента обработки данных, а именно языка программирования Python. Кроме того, сфера потребительских услуг никогда не потеряет свою актуальность что в 2021 году, что в 2031.

По результатам работы можно будет сделать вывод о взаимосвязи колебания логарифмической доходности индекса фондового рынка и колебания логарифмической доходности акций, входящих в его состав, значит, и их биржевых котировок.

# **I. ТЕОРЕТИЧЕСКАЯ СПРАВКА**

1. **Математическая статистика**

Часть единой дисциплины прикладной математики - теорией вероятностей и математической статистикой представляющая 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, называется математической статистикой.

Целью мат. статистики является получение прикладных и фундаментальных заключений путем создания единой методики по сбору и обработке [7].

1. **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, которые взаимоисключают друг друга, при этом назовём гипотезу H0 основной, а H1 – дополнительной. В дальнейшем принимаем в качестве базисного предположения утверждение о том, что одна из гипотез является справедливой.

Так как H0 и H1 – взаимоисключающие гипотезы, то принятие H0  автоматически означает отступление от гипотезы и наоборот отступление от гипотезы H0 означает принятие гипотезы H1.

Правило, по которому, если гипотеза H0 не принимается, а выборка принадлежит некоторой области К, и наоборот если гипотеза H0 принимается, а выборка не принадлежит некоторой области К, называется статистическим критерием или статистикой критерия. Область значений критерия, на которой принимается гипотеза H0, называют областью допустимых значений или же областью принятия гипотезы. Соответственно, область значений критерия, на которой отклоняется гипотеза H0 , называют критической областью.

Как правило критическую область можно задать, используя неравенства:

(2.1)

или

(2.2)

или

, (2.3)

где , - статистика критерия [9].

1. **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.

Уровень значимости критерия α - вероятность ошибки первого рода или значимость критерия. За мощность критерия обозначают 1-β, где β – вероятность ошибки второго рода.

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 вычислять соответствующее критическое значение [1].

1. **P-значения**

P\_value позволяет принять или отвергнуть нулевую гипотезу сразу для всех уровней значимости без необходимости вычислять критические значения.

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

Если представить, что P-value найдено или заранее известно, то принятие (или отклонение) Н0 для данного реализуется так: в случае, гипотеза H0 отклоняется при , и принимается при .

Рассмотрим отдельно случай, когда . Как правило, критическую область можно представить в виде

(4.1)

где – непрерывная убывающая функция. Как нетрудно видеть, в этом случае

)) (4.2)

и для имеет месторавенство означающее, что H0 принимается. Отсюда уже легко получитьшироко применяемую формулу:

(4.3)

Действительно, при любом уровне значимости

(4.4)

Аналогично , (4.5)

где – непрерывная возрастающая функция, Р-значение удовлетворяет отношению

(4.6) [1]

1. **Логарифмическая доходность**

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

(5.1)

Или же, применив свойство логарифма от частного, получим:

(5.2)

Где t – рассматриваемый период, Pt – цена акции в данный период, Pt-1 – 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% [3].

## **Критерий Колмогорова**

Также в данная работе будет использоваться вспомогательный критерий Колмогорова. Этот критерий позволяет подтвердить или опровергнуть гипотезу о распределении величины по какому-либо закону.

За статистику критерия Колмогорова принято считать меру расхождения между значением эмпирической функции распределения Fn(x) и предполагаемой теоретической функцией распределения F(x). Среди всех полученных значений выбирают максимальное по модулю

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:

. (6.1)

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

. (6.2)

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:

. (6.3)

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

, (6.4)

где

(6.5)

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.

Чтобы определить максимальное абсолютное отклонение гипотетической функции F(x) от эмпирической функ­ции применяется следующая формула:

, (6.6)

где - i-й член вариационного ряда

[1].

Таблица. Критические значения статистики критерия Колмогорова-Смирнова

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0,15 | 0,10 | 0,05 | 0,03 | 0,01 |
| DnH | 0,775 | 0,819 | 0,895 | 0,995 | 1,035 |

1. **Критерий Фишера**

Когда нужно сравнить две дисперсии нормальных распределений используют критерий Фишера. Принято считать, что все доходности финансовых активов подчиняются нормальному распределению по центральной предельной теореме: если количество дней, за которое рассчитывается логарифм кумулятивной доходности, является большим, то распределение суммы логарифмов стремится к нормальному. Но мы убедимся в этом по критерию Колмогорова.

Пусть имеются две независимые выборки из нормальных распределений:

Будем считать, что параметры известны. В качестве основной гипотезы примем H0: , а в качестве дополнительной одну из трёх гипотез:

1. H1 :
2. H1 :
3. H1 :

Теорема о построении критериев для проверки гипотезы c известным уравнением значимости :

Если верна Н0, то

где , , а F(m-1,n-1) – распределение Фишера с m-1 и n-1 степенями свободы [2].

# **II. ПРАКТИЧЕСКАЯ ЧАСТЬ**

1. **Предварительный анализ данных**

Начнём работу с предварительного анализа данных. Котировки компаний, используемые в работе, были взяты с сайта Межбанковского финансового дома [11]. Информацию об индексе и входящих в его состав акциях [10], а распределении акций по секторам [9]. Сопоставим котировки акций и название компаний в таблице 1.

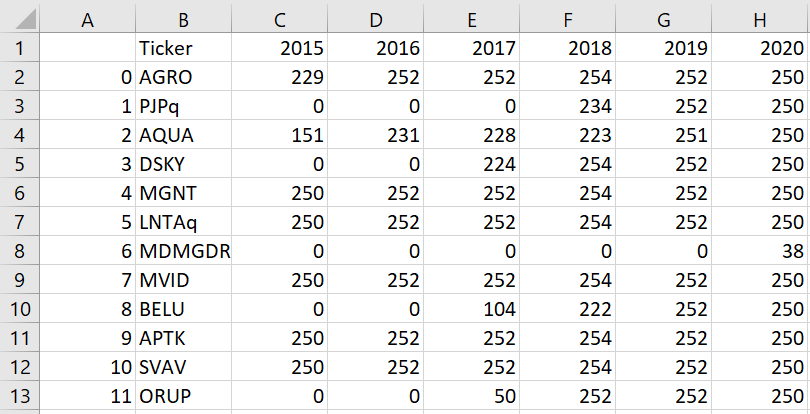
Таблица 1 - Список компаний и тикеров котировок акций

|  |  |
| --- | --- |
| Тикер | Название компании |
| AGRO | ПАО «РУСАГРО» |
| PJPq | X5 Retail Group |
| AQUA | ПАО «Русская аквакультура» |
| DSKY | ПАО «Детский мир» |
| MGNT | ПАО «Магнит» |
| LNTAq | ПАО "ЛЕНТА" |
| MDMGDR | МГП «Мать и Дитя» |
| MVID | ПАО «М.Видео» |
| BELU | ПАО «Белуга Групп» |
| APTK | ПАО «Аптечная сеть 36,6» |
| SVAV | ПАО «Со́ллерс» |
| ORUP | OR GROUP |

Для расчёта количества торговых дней использую столбец <DATE>, в которым указана информация о датах котировок. Она представлена в формате yyMMdd. Для вычисления торговых дней используется группировка данных по полю <DATE> с использованием информации о годе, после чего осуществляется подсчёт торговых дней в получившихся группах. Информация выводится в таблицу, где строчки – это названия компаний, а столбцы – года, и сохраняется в файл формата csv с разделителем «;».

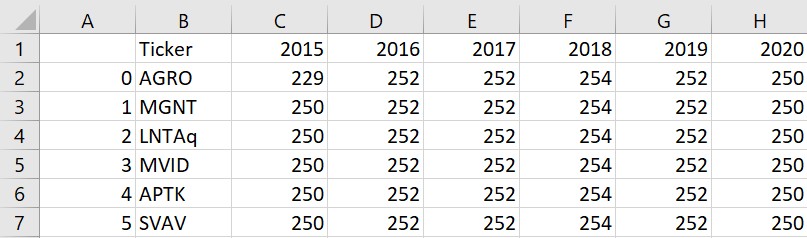
Полученные результаты количества торговых дней для всех компаний можем наблюдать в таблице 2.

Таблица 2 - Вычисленное количество торговых дней для всех компаний



Отметим, что у котировок PJPq данные с 2018 года, MDMGDR с 2020, у DSKY и BELU с 2017, у ORUP с конца 2017, а у AQUA с середины 2015, поэтому их анализировать дальше не будем. Составим новую таблицу дней, где будут только те компании, в которых есть торговые дни в период с 201 по 2020 года. Полученные результаты представлены в таблице 3.

Таблица 3 - Вычисленное количество торговых дней для сокращенного списка компаний (скорректированной таблицы)



Оставшиеся 6 компании имеют более 229 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. Для этого рассмотрим столбец <CLOSE>, в котором содержится цена, сформированная на момент закрытия акции. В таблице 4 мы можем наблюдать полученный результат для максимальных скачков цен вниз, а в таблице 5 – для максимальных скачков вверх. Для наглядности применим условное форматирование в MS Excel по цветам (зеленый – максимальное значение, красный – минимальное).

Таблица 4 - Максимальные относительные скачки цен вниз

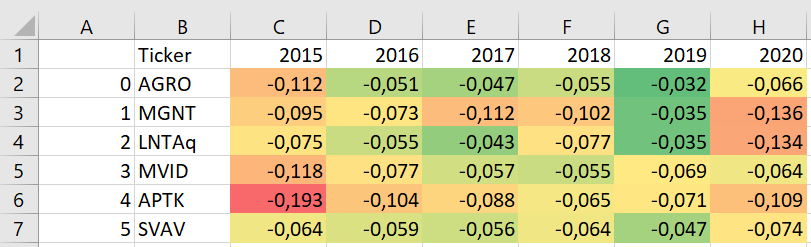
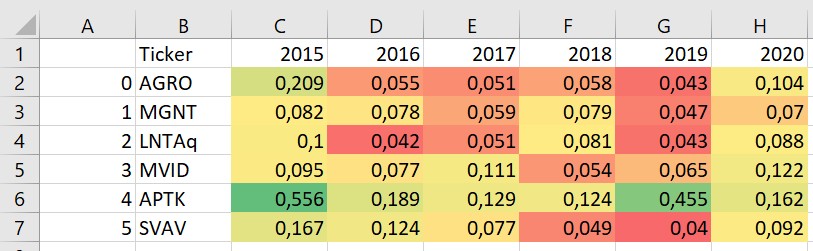


Таблица 5 - Максимальные относительные скачки цен вверх



Максимальный скачок вниз у APTK (19,3%) и у этой же компании самый сильный скачок вверх (55,6%), из чего можно сделать вывод, что ПАО «Аптечная сеть 36,6» обладает максимальными относительными изменениями цен. Построим график изменения цен для этой компании. Результаты отображены на рисунке 1.

Рисунок 1. График изменения цен для компании APTK

Так как скачок цены больше, чем 50%, значит данная компания не дает нам репрезентативную выборку, и мы исключаем её из дальнейшего исследования.

По итогам предварительного анализа данных мы будем рассматривать 5 компаний, которые имеют более 229 торговых дней в году и не имеют большого скачка цен, что предоставляет нам репрезентативную выборку, которую мы и будем исследовать.

## **2. Проверка гипотезы на модельных данных**

Проверим гипотезу о равенстве дисперсий логарифмической доходности фондового индекса и входящих в его состав акций для модельных данных, сгенерированных случайным образом. Используя метод Монте-Карло, сгенерируем две выборки из 252 элементов с помощью критерия Колмогорова, чтобы проверить гипотезу на модельных данных, что позволит убедиться в корректной работе программы, вычисляющей значение критерия и p-значения.

Также построим гистограмму P-значений для смоделированных. Она представленна на рисунке 2 и 3.

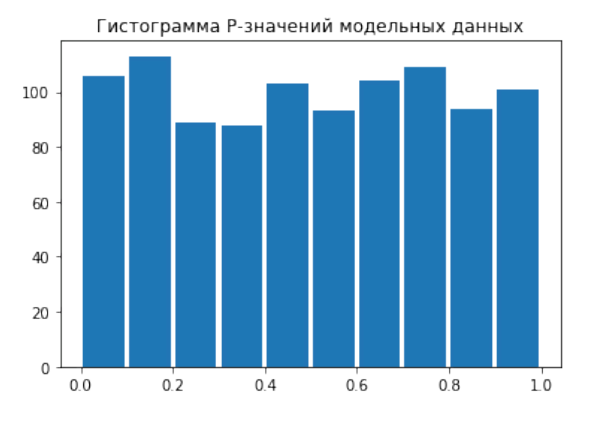


Рисунок 2. Гистограмма P-значения для модельных данных 1

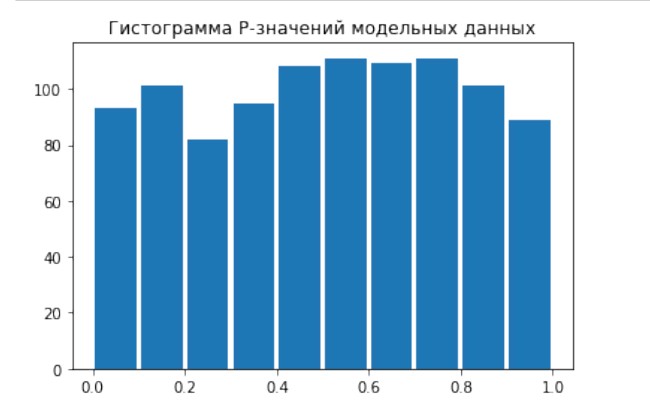


Рисунок 3. Гистограмма P-значения для модельных данных 2

Получили P-значение критерия Колмогорова равным 0.7810093429954852 для первой выборки и 0.9155674384900221 для второй выборки, что говорит о равномерности распределений P-значений для модельных данных [8].

Проверим Критерий Фишера на модельных данных. Критерий Фишера равен 1.2487322294814869, значит дисперсии равны и можно сделать вывод, что гипотеза о равенстве дисперсий логарифмической доходности фондового рынка и входящих в его состав акций принимается.

## **3. Проверка гипотезы на реальных данных**

Убедившись в том, что все необходимые нам программы работают корректно и обработка данных происходит корректно, проверим выполнение гипотезы о равенстве дисперсий логарифмической доходности фондового рынка и входящих в его состав акций по критерию Фишера на реальных данных, а именно на котировках акций компаний, которых входят в индекс ММВБ потребительский серктор. Как и на этапе предварительной обработке данных возспользуемся полями «<DATE>», показывающее дату информации о котировках, и «<CLOSE>», показывающее цену последней сделки, которая была совершена. Отметим, что критерий работает с логарифмической доходностью, которой в исходных данных нет, соответственно вычислим её и сохраним значение в списке. Для этого задана специальная функция.

После чего вычислим P-значения уже для реальных данных, что тоже делается через заранее прописанную функцию, для выборки за все 5 лет по каждому году, с помощью Критерия Колмогорова. После чего выведем полученный результат, которая представлен в таблице 6.

Таблица 6. P-значений для реальных данных за 5 лет по каждому году



Отразим полученные значения на гистограмме.

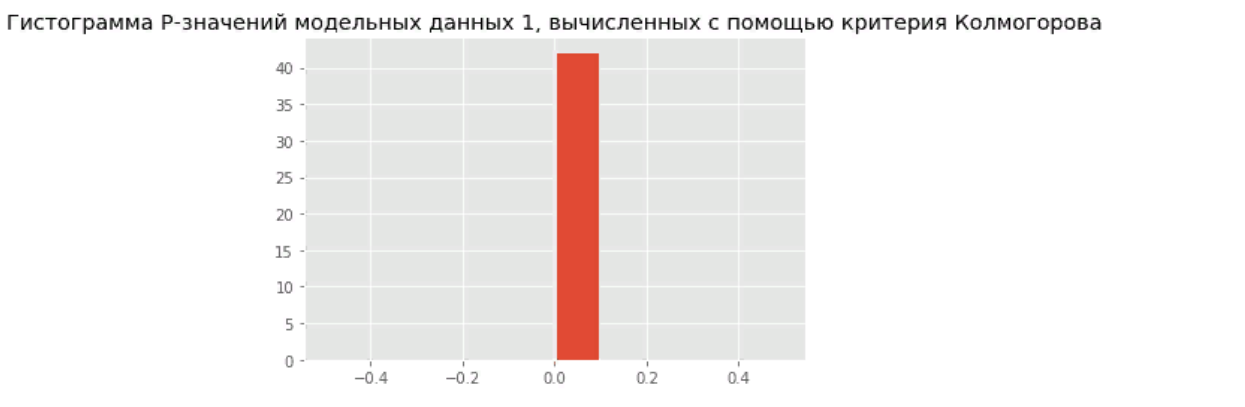


Рисунок 4. Гистограмма P-значения.

Так как значения P-значения сосредотачиваются вокруг нуля, то мы не можем сказать, что реальные данные распределены равномерно.

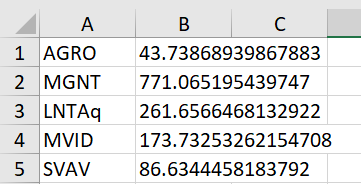
Вычислим значения дисперсий для каждой компании по каждому году и получим следующие значения.

Таблица 7. Дисперсии логарифмической доходности за 5 лет по каждому году



В дисперсионном анализе не применяется прямое сравнение дисперсий, поэтому переходим к анализу значений логарифмической доходности, но уже с применением рассматриваемого критерия Фишера. Для этого воспользуемся специально написанной для этого функцией для компаний и индекса MOEXFN и получим значения, которые представлены в таблице 8.

Таблица 8. Проверка критерия Фишера



У логарифмической доходности акций MGNT было наибольшее значение дисперсии, для неё мы получили наибольшее значение критерия Фишера. У логарифмической доходности акций AGRO было наименьшее значение дисперсии, для неё мы получили наименьшее значение критерия Фишера.

Для нормально распределённой выборки на модельных данных мы получили значение Фишера. По реальным данным видно, что эти значения сильно отличаются от нормально распределённой.

Значит, можем сделать вывод, что гипотеза о равенстве дисперсий логарифмической доходности индекса фондового рынка и входящих в его состав акций отвергается.

## **4. Альтернативные гипотезы и оценка мощности критерия**

В качестве альтернативных гипотез было предположено, что логарифмическая доходность имеет:

1. распределение по модулю закона Стьюдента с тремя степенями свободы;
2. логнормальное распределение.

Вычислим мощность критерия Фишера, вычисляя 1000 раз Р-значения при уровне значимости для каждого альтернативного распределения.

Результаты можно видеть в таблице 15.

Таблица 9. Мощность критерия Фишера для различных распределений.

|  |  |  |  |
| --- | --- | --- | --- |
| Вид распределения | 20 | 125 | 252 |
| Распределение по модулю закона Стьюдента с тремя степенями свободы | 0.6406 | 0.9874 | 0.9998 |
| Логнормальное распределение со стандартным отклонением равным 1/4 | 0.6603 | 0.9893 | 0.9989 |

Из таблицы следует, что мощность критерия Фишера для логнормального распределения и для распределения Стьюдента с тремя степенями свободы увеличивается с увеличением объёма выборки. Значит, чем больше выборка, тем меньше вероятность допустить ошибку второго рода.

# **ЗАКЛЮЧЕНИЕ**

В данной курсовой работе была рассмотрена проверка гипотезы о равенстве дисперсий логарифмической доходности индекса фондового рынка и входящих в его состав акций по критерию Фишера на примере котировок акций компаний, входящих в индекс MOEXCN (потребительского сектора), а также рассмотрена правильность выполнения программ как на модельных данных, так и на реальных.

В результате проверки поставленной гипотезы о равенстве дисперсий логарифмической доходности индекса фондового рынка и входящих в его состав акций я получила следующие результаты: гипотеза не подтвердилась для выбранных мной реальных данных, но зато подтвердилась для сгенерированных случайным образом модельных данных.

Из полученных результатов следует лишь одно: выбранная гипотеза не верна для реальных данных, а значит дисперсии логарифмической доходности индекса фондового рынка и входящих в его состав акций не равны. На подобный результат могло повлиять множество факторов. Одним из них можно назвать и то, что у каждого из критериев существуют недостатки. Так, например, критерий Фишера крайне чувствителен к отклонению от нормального распределения и достигает максимальной эффективности для выборок одинакового объёма, что крайне редко можно наблюдать в современных реалиях, а не на моделях.

Но сам факт того, что на модельных данных, сформированных случайным образом из диапазона чисел [0;1], гипотеза подтверждается показывает то, что в вычислениях нет ошибок, а значит и критерии были проверены правильно.

# **СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ**

1. А.В. Браилов, А.С. Солодовников; под ред. В.А. Бабайцева, В.Б. Гисина Математика в экономике: учебник: В 3-х ч. Ч. 3. Теория вероятностей и математическая статистика – М.: Финансы и статистика, 2013
2. Браилов А.В. Лекции по математической статистике. – М.: Финакадемия, 2008.
3. Браилов А.В. Лекции по теории вероятности. – М.: Финакадемия, 2008.
4. Красс М. С., Б.П. Чупрынов Б.П. Математика в экономике - М.: Финансы и статистика, 2007.
5. С.С. Бондарчук, И.С. Бондарчук Статобработка экспериментальных данных в MS Excel: учебное пособие. Издательство Томского государственного педагогического университета, 2018.
6. Глебов Криволапов Практикум по математической статистике. Проверка гипотез с использованием Excel, MatCale, R и Python. М.: Прометей, 2019.
7. Малугин В.А. Математическая статистика. М.: Юрайт, 2020.
8. В.Я. Крохалев, С.А. Скопинов, В.А. Телешев СТАТИСТИКА, Издательство УГМУ 2018
9. Смирнова З.М., Крейнина М.В. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ.
10. <https://smart-lab.ru/q/index_stocks/MOEXCN/>
11. <https://ru.investing.com/equities/>
12. <https://mfd.ru/>

# **ПРИЛОЖЕНИЯ**

## Приложение 1

Технические характеристики компьютера:

Процессор IntelI CoreI 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

Время выполнения программы:

--- 26.84362006187439 seconds ---

## Приложение 2

Список файлов

|  |
| --- |
| Имя файла |
| Таблица 1 – Список компаний и тикеров котировок акций |
| Таблица 2 - Вычисленное количество торговых дней для всех компаний |
| Таблица 3 - Вычисленное количество торговых дней для сокращенного списка компаний |
| Таблица 4 - Максимальные относительные скачки цен вниз |
| Таблица 5 - Максимальные относительные скачки цен вверх |
| Рисунок 1. График изменения цен для компании APTK |
| Рисунок 2. Гистограмма P-значений для модельных данных 1 |
| Рисунок 3. Гистограмма P-значений для модельных данных 2 |
| Таблица 6. P-значений для реальных данных за 5 лет по каждому году |
| Рисунок 4. Гистограмма P-значения. |
| Таблица 7. Дисперсии логарифмической доходности за 5 лет по каждому году |
| Таблица 8. Проверка критерия Фишера |
| Таблица 9. Мощность критерия Фишера для различных распределений. |

## Приложение 3

import time

start\_time = time.time() # время начала

#загружаем необходимые библиотеки

import numpy as np

import pandas as pd

import statistics as st

import scipy.stats as stats

import matplotlib.pyplot as plt

from scipy.stats import norm

import csv

from math import \*

from scipy.stats import chisquare

import matplotlib

import math

**Таблица 2 - Вычисленное количество торговых дней для всех компаний**

#посчитаем количество торговых дней для всех компаний

tickers=['AGRO','PJPq', 'AQUA', 'DSKY', 'MGNT', 'LNTAq', 'MDMGDR', 'MVID', 'BELU', 'APTK', 'SVAV', 'ORUP']

kolvo\_dney\_vse=pd.DataFrame()

kolvo\_dney\_vse['Ticker'] = tickers

years=range(2015,2021)

for i in years:

kolvo\_year=[]

for j in range(len(tickers)):

file=pd.read\_csv(tickers[j]+'.csv',sep=';')

diap=(file['<DATE>']>=i\*10000)&(file['<DATE>']<(i+1)\*10000)

kolvo=len(file[diap])

kolvo\_year.append(kolvo)

kolvo\_dney\_vse[str(i)]=kolvo\_year

kolvo\_dney\_vse.to\_csv('Количество торговых дней всех компаний по годам.csv',sep=';')

kolvo\_dney\_vse

**Таблица 3 - Вычисленное количество торговых дней для сокращенного списка компаний**

#посчитаем количество торговых дней компаний для скорректированной таблицы

tickers=['AGRO', 'MGNT', 'LNTAq', 'MVID', 'APTK', 'SVAV']

kolvo\_dney\_kor=pd.DataFrame()

kolvo\_dney\_kor['Ticker'] = tickers

years=range(2015,2021)

for i in years:

kolvo\_year=[]

for j in range(len(tickers)):

file=pd.read\_csv(tickers[j]+'.csv',sep=';')

diap=(file['<DATE>']>=i\*10000)&(file['<DATE>']<(i+1)\*10000)

kolvo=len(file[diap])

kolvo\_year.append(kolvo)

kolvo\_dney\_kor[str(i)]=kolvo\_year

kolvo\_dney\_kor.to\_csv('Количество торговых дней компаний для скорректированной таблицы.csv',sep=';')

kolvo\_dney\_kor

**Таблица 4 - Максимальные относительные скачки цен вниз**

#посчитаем относительные скачки цен вниз

tickers=['AGRO', 'MGNT', 'LNTAq', 'MVID', 'APTK', 'SVAV']

down=pd.DataFrame()

down['Ticker'] = tickers

years=range(2015,2021)

for i in years:

down\_znach=[]

for j in range(len(tickers)):

file=pd.read\_csv(tickers[j]+'.csv',sep=';')

file['<LEAP>'] = file['<CLOSE>'].pct\_change().round(3)

diap=(file['<DATE>']>=i\*10000)&(file['<DATE>']<(i+1)\*10000)

znach\_year=file[diap]['<LEAP>'].min()

down\_znach.append(znach\_year)

down[str(i)]=down\_znach

down.to\_csv('Относительные изменения цен вниз.csv',sep=';', decimal=',')

down

**Таблица 5 - Максимальные относительные скачки цен вверх**

#посчитаем относительные скачки цен вверх

tickers=['AGRO', 'MGNT', 'LNTAq', 'MVID', 'APTK', 'SVAV']

up=pd.DataFrame()

up['Ticker'] = tickers

years=range(2015,2021)

for i in years:

up\_znach=[]

for j in range(len(tickers)):

file=pd.read\_csv(tickers[j]+'.csv',sep=';')

file['<LEAP>'] = file['<CLOSE>'].pct\_change().round(3)

diap=(file['<DATE>']>=i\*10000)&(file['<DATE>']<(i+1)\*10000)

znach\_year=file[diap]['<LEAP>'].max()

up\_znach.append(znach\_year)

up[str(i)]=up\_znach

up.to\_csv('Относительные изменения цен вверх.csv',sep=';', decimal=',')

up

#проверка гипотезы на модельных данных

kvanlil\_1 = [kv for kv in np.arange(0.1,1,0.1)]

znach\_1=[]

for i in kvanlil\_1:

znach\_1.append(np.quantile(Shapir1,i))

df\_1 = pd.DataFrame({'Квантиль':kvanlil\_1, 'Значение':znach\_1})

df\_1.to\_csv('Первые 9 квантилей статистики Шапиро-Уилка1.csv',sep=';')

df\_1

kvanlil\_2 = [kv for kv in np.arange(0.1,1,0.1)]

znach\_2=[]

for i in kvanlil\_2:

znach\_2.append(np.quantile(Shapir2,i))

df\_2 = pd.DataFrame({'Квантиль':kvanlil\_2, 'Значение':znach\_2})

df\_2.to\_csv('Первые 9 квантилей статистики Шапиро-Уилка2.csv',sep=';')

df\_2

**Таблица 6 - Таблица квантилей 1**

from scipy.stats import kstest

kvanlil\_9991 = [kv for kv in np.arange(0.001,1,0.001)]

znach\_9991 = []

for i in kvanlil\_9991:

znach\_9991.append(np.quantile(Shapir1,i))

print(kstest(znach\_9991,'norm'))

df\_9991 = pd.DataFrame({'Квантиль':kvanlil\_9991, 'Значение':znach\_9991})

df\_9991.to\_csv('999 квантилей статистики Шапиро-Уилка1.csv')

df\_9991

**Рисунок 2. Гистограмма P-значений для модельных данных 1**

pvalues=[]

for i in range(1000):

x=np.random.normal(loc=0, scale=1, size=252)

pvalues.append(stats.shapiro(x)[1])

plt.hist(pvalues,rwidth=0.9)

plt.title('Гистограмма Р-значений модельных данных')

plt.show()

**Таблица 7 - Таблица квантилей 2**

from scipy.stats import kstest

kvanlil\_9992 = [kv for kv in np.arange(0.001,1,0.001)]

znach\_9992 = []

for i in kvanlil\_9992:

znach\_9992.append(np.quantile(Shapir2,i))

print(kstest(znach\_9992,'norm'))

df\_9992 = pd.DataFrame({'Квантиль':kvanlil\_9992, 'Значение':znach\_9992})

df\_9992.to\_csv('999 квантилей статистики Шапиро-Уилка2.csv')

df\_9992

**Рисунок 3. Гистограмма P-значений для модельных данных 2**

pvalues1=[]

for i in range(1000):

x=np.random.normal(loc=0, scale=1, size=252)

pvalues1.append(stats.shapiro(x)[1])

plt.hist(pvalues1,rwidth=0.9)

plt.title('Гистограмма Р-значений модельных данных')

plt.show()

#выполним проверку с помощью критерия Колмогорова

def p\_kolm(n):

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

p=stats.kstest(viborka,'norm')

return round(p[1],3)

#построение гистограммы Р-значений, вычисленных с помощью критерия Колмогорова

from scipy.stats import kstest

pvalues\_kolm1=[]

for i in range(1000):

pvalues\_kolm1.append(p\_kolm(n))

print(kstest(pvalues\_kolm1,'norm'))

plt.hist(pvalues\_kolm1,rwidth=0.9)

plt.title('Гистограмма Р-значений модельных данных 1, вычисленных с помощью критерия Колмогорова')

plt.show()

#построение гистограммы Р-значений, вычисленных с помощью критерия Колмогорова

from scipy.stats import kstest

pvalues\_kolm2=[]

for i in range(1000):

pvalues\_kolm2.append(p\_kolm(n))

print(kstest(pvalues\_kolm2,'norm'))

plt.hist(pvalues\_kolm2,rwidth=0.9)

plt.title('Гистограмма Р-значений модельных данных 1, вычисленных с помощью критерия Колмогорова')

plt.show()

# проверим, что наши выборки взяты из распределения одного вида

viborka\_1=np.random.normal(loc=0, scale=1, size=n)

viborka\_2=np.random.normal(loc=0, scale=1, size=n)

pvalue1 = stats.ks\_2samp(viborka\_1,viborka\_2)

print(pvalue1[1])

#проверка притерия Фишера

def fisher\_criterion(v1, v2):

return (abs(np.mean(v1) - np.mean(v2)))/ (np.var(v1) + np.var(v2))

fisher\_criterion (znach\_9991, znach\_9992)

#проверка мощьности критерия

m = 10000 #количество экспериментов

f1=[]

f2=[]

for i in range(m):

viborka11 = stats.t.rvs(df=3, loc = 0, scale = 1,size = 252)

viborka12 = stats.t.rvs(df=3, loc = 0, scale = 1,size = 252)

viborka21 = stats.lognorm.rvs(s=1/4 ,loc = 0, scale = 1,size = 252)

viborka22 = stats.lognorm.rvs(s=1/4 ,loc = 0, scale = 1,size = 252)

f1.append(fisher\_criterion(viborka11, viborka12))

f2.append(fisher\_criterion(viborka21, viborka22))

pow1 = 0

pow2 = 0

**Таблица 11. Мощность критерия Фишера для различных распределений.**

## Вычисление мощности критерия

for i in f1:

if i<0.05:

pow1+=1

print(pow1/m)

for i in f2:

if i<0.05:

pow2+=1

print(pow2/m)

#проверка гипотезы на реальных данных

#импорт таблиц

tickers=['AGRO', 'MGNT', 'LNTAq', 'MVID', 'SVAV', 'MOEXCN']

def tables(ticker):

company = pd.read\_csv(ticker+'.csv', sep=';') #чтение из файла

company['<DATE>'] = pd.to\_datetime(company['<DATE>'], dayfirst=True, infer\_datetime\_format=True) #преобразование даты

return company

# список логарифмов доходности по годам

def logYeld(ticker, a):

LD = []

ya=['2015', '2016', '2017', '2018', '2019', '2020']

company = tables(ticker)

company['Логарифм доходности'] = np.log(company['<CLOSE>'])

for year in ya:

table = pd.read\_csv('Количество торговых дней компаний для скорректированной таблицы.csv', sep=';')

n = (table[str(year)][a])

z = company['Логарифм доходности'][company['<DATE>']<=str(year)+'-12-31'].size-(n+1)

LD.append((company['Логарифм доходности'][company['<DATE>']<=str(year)+'-12-31'])[z:].kurtosis())

print (LD)

return LD

matplotlib.style.use('ggplot')

PV = []

for ticker in tickers:

a = tickers.index(ticker)

log = logYeld(ticker, a)

**Таблица 8. P-значений для реальных данных за 5 лет по каждому году**

def log\_prof\_max(year,file): #cоздаю функцию для подсчета логарифмической доходности

company = pd.read\_csv(file, sep=';') #считываю файл

#обрезаю дату до четырех цифр года

company['<DATE>'] = [int(str(company['<DATE>'][i])[:4]) for i in range(len(company['<DATE>']))]

number = list(np.where(company['<DATE>'] == year)[0])[:-1]

#ищу логарифмическую доходность

#округляю до двух знаков после запятой

logg\_doch\_max = np.log(company['<CLOSE>'])

return stats.kstest(logg\_doch\_max, 'norm')[1]

#далее записываю в массив названия тикеров компаний для формирования столбца DataFrame и считывания файлов

tickers=['AGRO', 'MGNT', 'LNTAq', 'MVID', 'APTK', 'SVAV', 'MOEXCN']

#для формирования названий столбцов DataFrame создаю переменную, хранящую массив всех годов с 2015 по 2020

years = range(2015,2021)

#создаю DataFrame

table\_log\_max = pd.DataFrame() #добавляю номер тикера с 1 для наглядности

#записываю в столбец названия тикеров

table\_log\_max['Тикер'] = tickers

#высчитываю максимальное отклонение цен по каждому году

all\_val=[]

for year in years:

log\_number\_max = [] #создаю пустой массив для заполнения максимального отклонения по каждой акции

for ticker in tickers:

#добавляю в массив значение логарифмических доходностей

#по каждому году исследуемой акции из файла рассчитываю функции и добавляю в массив

log\_number\_max.append(round(log\_prof\_max(year, ticker+'.csv'), 10))

all\_val.append(round(log\_prof\_max(year, ticker+'.csv'), 10))

#записываю в столбец по году значения массива log\_number\_max, который на каждом шаге цикла

#хранит логарифмическую доходность в определенный год по каждой акции

table\_log\_max[str(year)] = log\_number\_max

#текст вывожу жирным

print('\033[1m' + 'Таблица. Р-значения по критерию Колмогорова')

table\_log\_max

plt.hist(all\_val,rwidth=0.9)

plt.title('Гистограмма Р-значений модельных данных 1, вычисленных с помощью критерия Колмогорова')

plt.show()

**Таблица 9. Дисперсии логарифмической доходности за 5 лет по каждому году**

def log\_prof\_max(year,file): #cоздаю функцию для подсчета логарифмической доходности

company = pd.read\_csv(file, sep=';') #считываю файл

#обрезаю дату до четырех цифр года

company['<DATE>'] = [int(str(company['<DATE>'][i])[:4]) for i in range(len(company['<DATE>']))]

number = list(np.where(company['<DATE>'] == year)[0])[:-1]

#ищу логарифмическую доходность (т.е логарифм отношения цены закрытия текущей акции к предыдущей)

#округляю до двух знаков после запятой

logg\_doch\_max = np.log(company['<CLOSE>'])

return np.var(logg\_doch\_max)

#далее записываю в массив названия тикеров компаний для формирования столбца DataFrame и считывания файлов

tickers=['AGRO', 'MGNT', 'LNTAq', 'MVID', 'SVAV', 'MOEXCN']

#для формирования названий столбцов DataFrame создаю переменную, хранящую массив всех годов с 2015 по 2020

years = range(2015,2021)

#создаю DataFrame

table\_log\_max = pd.DataFrame() #добавляю номер тикера с 1 для наглядности

#записываю в столбец названия тикеров

table\_log\_max['Тикер'] = tickers

#высчитываю максимальное отклонение цен по каждому году

for year in years:

log\_number\_max = [] #создаю пустой массив для заполнения максимального отклонения по каждой акции

for ticker in tickers:

#добавляю в массив значение логарифмических доходностей

#по каждому году исследуемой акции из файла рассчитываю функции и добавляю в массив

log\_number\_max.append(log\_prof\_max(year, ticker+'.csv'))

#записываю в столбец по году значения массива log\_number\_max, который на каждом шаге цикла

#хранит максимальную логарифмическую доходность в определенный год по каждой акции

table\_log\_max[str(year)] = log\_number\_max

#текст вывожу жирным

print('\033[1m' + 'Таблица. Дисперсии логарифмической доходности')

table\_log\_max

**Таблица 10. Проверка критерия Фишера**

def log\_prof\_max(file): #cоздаю функцию для подсчета логарифмической доходности

company = pd.read\_csv(file, sep=';') #считываю файл

#обрезаю дату до четырех цифр года

company['<DATE>'] = [int(str(company['<DATE>'][i])[:4]) for i in range(len(company['<DATE>']))]

#ищу логарифмическую доходность (т.е логарифм отношения цены закрытия текущей акции к предыдущей)

#округляю до двух знаков после запятой

logg\_doch\_max = np.log(company['<CLOSE>'])

return (logg\_doch\_max)

#далее записываю в массив названия тикеров компаний для формирования столбца DataFrame и считывания файлов

tickers=['AGRO', 'MGNT', 'LNTAq', 'MVID', 'SVAV', 'MOEXCN']

#для формирования названий столбцов DataFrame создаю переменную, хранящую массив всех годов с 2015 по 2020

#создаю DataFrame

table\_log\_max = pd.DataFrame() #добавляю номер тикера с 1 для наглядности

#записываю в столбец названия тикеров

table\_log\_max['Тикер'] = tickers

#высчитываю максимальное отклонение цен по каждому году

log\_number\_max = [] #создаю пустой массив для заполнения максимального отклонения по каждой акции

for ticker in tickers:

#добавляю в массив значение логарифмических доходностей

#по каждому году исследуемой акции из файла рассчитываю функции и добавляю в массив

log\_number\_max.append(log\_prof\_max(ticker+'.csv'))

#записываю в столбец по году значения массива log\_number\_max, который на каждом шаге цикла

#хранит максимальную логарифмическую доходность в определенный год по каждой акции

table\_log\_max = log\_number\_max

#текст вывожу жирным

print('\033[1m' + 'Таблица. Логарифмическая доходность:')

table\_log\_max

for i in table\_log\_max[:5]:

y = fisher\_criterion(i, table\_log\_max[5])

print(y)

finish\_time = time.time()

print("--- %s seconds ---" % (finish\_time - start\_time)) # Время выполнения программы в секундах