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:

[Коровин Д.И](javascript:;).

Москва 2021

Оглавление

Введение**1**

I. Предварительный анализ данных 2

II. Теоретическая справка3

1.Математическая статистика……………………………………………………4

2.Статистическая гипотеза**5**

3.Ошибки первого и второго рода…………………………………………………...6

4.Статистический критерий7

5.P-значение (P-value) статистического критерия8

6.Критерии Хегази-Грина.................................................................................….9

7.Логарифмическая доходность………………………………………………...10

III. Практическая часть .........................................................................................14

1. Выбор альтернативной гипотезы и оценка мощности критерия ..........................16

2.Проверка гипотез для реальных данных.........................................................13

Заключение ...........................................................................................................14

Список используемых источников.....................................................................15

Приложения .................................................................................................................... 21

Приложение 1 ………………………………….............................................................. 21

Технические характеристики компьютера ………................................................ .... 21

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

Приложение 2 ………………........................................................................................ 22

Коды программ ………………....................................................................................... 22

Приложение 3 ……………….......................................................................................... 41

Список файлов ………………....................................................................................... 41

**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 использования нейросети нужно подавать на вход нормально распределённые данные. В курсовой работе мы будем проверять, является ли всега логарифмическая доходность акций нормально распределенной переменно

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.

**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.

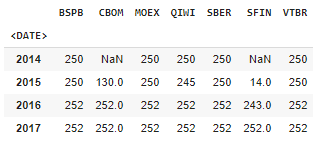
Таблица 1 - Список компаний и тикеров котировок акций

|  |  |
| --- | --- |
| 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.

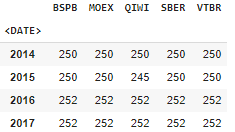
Проведем расчет количества торговых дней с последующим выводом результатов в таблицу, где по столбцам названия компании, а по строчкам – года. Для этого будет использовать поле DATE, показывающее дату информации о котировках. Оно представлено в формате дд/мм/гг. Для вычисления торговых дней используется 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.



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.

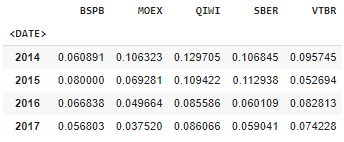


Оставшиеся 4 компании имеют 248 и более торговых дней в каждом из рассматриваемых годов, что обеспечивает достаточное количество информации для дальнейшего анализа и обработки.

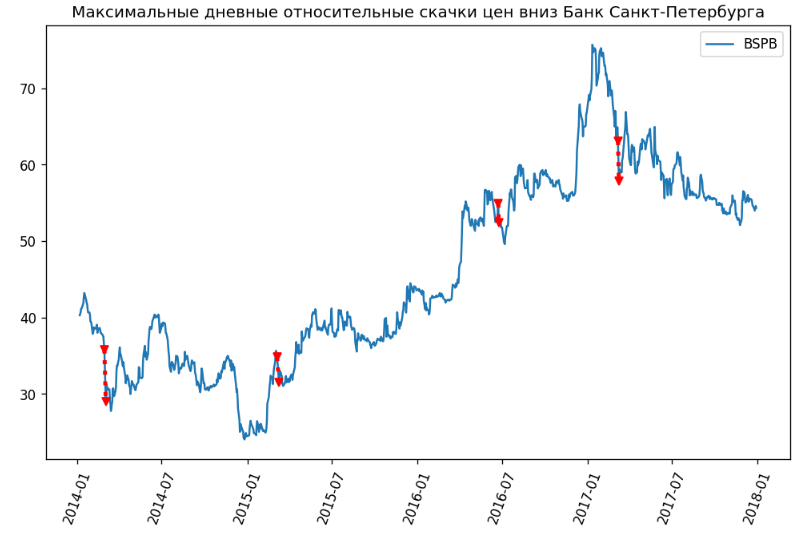
Далее рассмотрим максимальные относительные скачки цен вверх и вниз. Нам потребуется поле <CLOSE>, показывающее цену последней сделки, которая была совершена. В таблице 3 мы можем наблюдать полученный результат для максимальных скачков цен вниз, а в таблице 4 – для максимальных скачков вверх.

Таблица 4 - Вниз максимальные относительные скачки цен

Таблица 5 - Вверх максимальные относительные скачки цен



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.



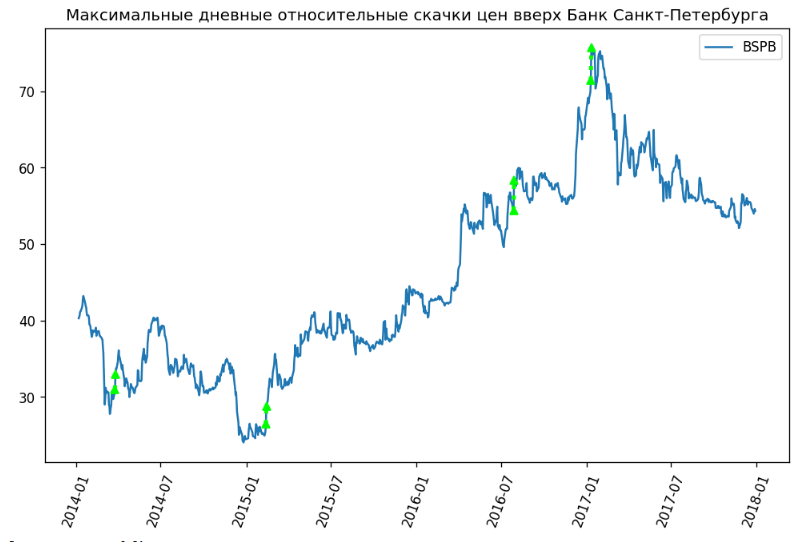
Рисунок 1. График изменения цен для компании BSPB максимальные скачки вниз

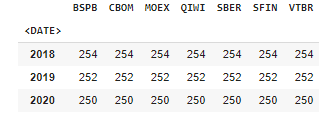
Рисунок 2. График изменения цен для компании BSPB максимальные скачки вверх

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.

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

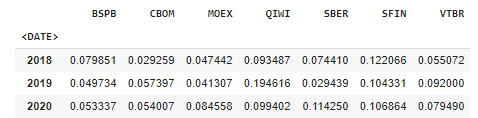


Таблица 8 - Вниз максимальные относительные скачки цен.



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.

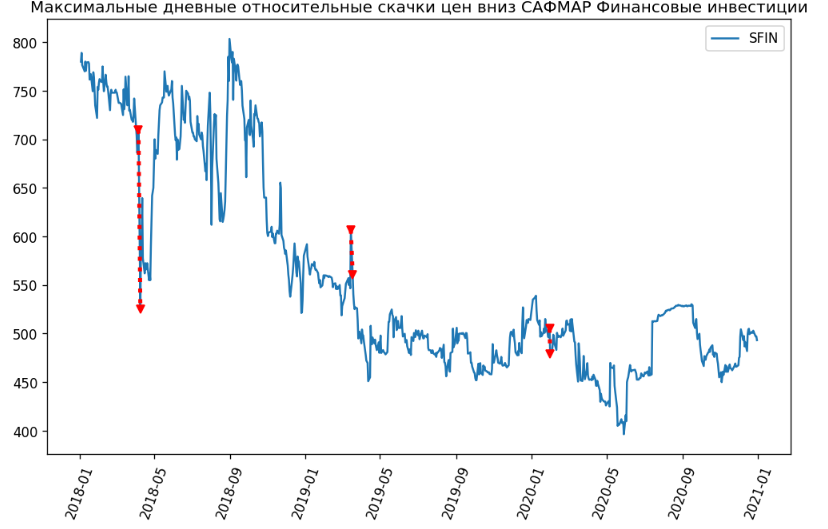


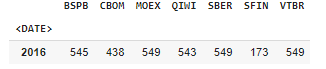
Рисунок 3. График изменения цен для компании SFIN максимальные скачки вниз



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

Следующий период первый квартал 2016 года с таймфреймом час.

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



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 называют гипотезу о виде неизвестного распределения или о параметрах известного распределения. Например, гипотеза H0 - случайная величина распределена по нормальному закону. Нулевой (основной) называется выдвинутая гипотеза H0. Альтернативной (конкурирующей) называется гипотеза, противоречащая основной (конкурирующих гипотез может быть несколько).

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:

K=x1,...,xnRn:t>c

 (1.1)

или

K=x1,...,xnRn:t<c

(1.2)

или

K=x1,...,xnRn:t<c1x1,...,xnRn:t>c2,

(1.3)

где c,c1,c2(c2>c1)=const , t=t(x1,...,xn) - 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 вычислять соответствующее критическое значение.

**Статистический критерий**

Статистическим критерием (К) называется случайная величина, точное или приближённое распределение, которой известно и которая служит для проверки справедливости нулевой гипотезы.

Множество возможных значений критерия делится на две непересекающихся области:

1) значения, при которых нулевая гипотеза справедлива (область принятия гипотезы).

2) значения, при которых нулевая гипотеза отвергается (критическая область).

Критическая область может быть односторонней (левосторонней, правосторонней) или двусторонней.

Точка Ккр, отделяющая критическую область от области принятия гипотезы, называется критической точкой. Чтобы определить критическую область, выбирают число q-уровень значимости. q- вероятность того, что при справедливости нулевой гипотезы значение критерия К попадает в критическую область. Тогда для правосторонней критической области Ккр определяется из условия:  P { K> Kkp } = q. Значение критерия табулировано, т. е. Kkp можно найти по таблице распределения критических точек в зависимости от уровня значимости q и числа степеней свободы f. -Наблюдаемое значение критерия Kнабл определяется по результатам эксперимента. Если Kнабл<Kkp, то гипотеза H0 принимается. Если Kнабл>Kkp, то H0 отвергается, а принимается конкурирующая гипотеза H1.

Для левосторонней критической области критическая точка определяется из условия:  P { K< Kkp } = q. Для двухсторонней:  P { K< K’kp } + P { K> K”kp } = q.

Если двусторонняя область симметрична относительно начала координат, то: P { K< K’kp } = q/2. Так как наблюдаемое значение критерия определялось по результатам эксперимента, то Кнабл - случайная величина и, следовательно, могут возникать ошибки при принятии гипотезы. Различают ошибки первого и второго рода. К ошибкам первого рода относят те, при которых отвергается правильная гипотеза. К ошибкам второго рода, относят те, при которых принимается неправильная гипотеза. Допустимой вероятностью ошибки первого рода является q-уровень значимости. Однако. если уменьшать q, то возрастает вероятность принятия неверной гипотезы, т. е. вероятность ошибок второго рода. Если справедлива гипотеза H1, то это считается доказанным, если справедлива гипотеза H0-то говорят, что результаты эксперимента не противоречат нулевой гипотезы. Для того чтобы считать H0 доказанной нужно или вновь повторить эксперимент или проверить гипотезу с помощью других критериев.

**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 выборки X=(X1, …, Xn) называется такое число PV(x)  , что  PV(x)≥α для любого уровня значимости α, при котором гипотеза H0 принимается, и PV(x)≤α, 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 PV(x)<α, гипотеза H0 отвергается, а если PVx>α 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 (PV<α) равна вероятности получить большое P-значение (PV>1-α). Однако, если H0 не верна, наблюдаемые P-значения (при достаточно высокой мощности критерия) концентрируются около нуля.

Р - значение находится из равенства: P-value =  PH0(f(x) > fнаблюдения).

Вспомогательный критерий для проверки статистической гипотезы, Критерий Колмогорова.

Используется по следующему алгоритму:

1. Строятся эмпирическая функция распределения Fn(x) и предполагаемая  теоретическая функция распределения F(x).

2. Определяется мера расхождения между теоретическим и эмпирическим распределением D по формуле D = max |Fn(x) - F(x)| и вычисляется величина.  λ=D√n.

3. Если вычисленное значение λ окажется больше критического λα, определенного на уровне значимости α, то нулевая гипотеза Н0 о том, что случайная величина имеет заданный закон распределения, отвергается. Если λ≤ λα, то считают, что гипотеза Н0 не противоречит опытным данным.

**Критерии Хегази-Грина**

В данной работе в качестве основного статистического критерия применяем критерии Хегази-Грина.  Хегази и Грин предложили  критерии со статистиками:

T1 =1ni=1n│Zi-ηi│

T2 =1ni=1n{Zi-ηi}2

Где zi=xi-xS,         x=1ni=1nxi,        S2=1n-1i-1n(xi-x)2 ,

X1…Xn – выборка объема n, распределённая по некоторому закону L(X), из некоторого генерального распределения. i *–* математическое ожидание i-й порядковой статистики стандартного нормального закона, которое можно найти из соотношения i=Ф-1in+1. Подчеркнем, что в статистике должна использоваться именно несмещенная оценка дисперсии. Распределения этих статистик очень сильно зависят от объема выборки.

**Логарифмическая доходность**

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

ln PtPt-1

(3.1)

Или же, применив свойство логарифма от частного, получим:

ln Pt- ln Pt-1

(3.2)

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%.

**III. ПРАКТИЧЕСКАЯ ЧАСТЬ**

1. Выбор альтернативной гипотезы и оценка мощности критерия

Дальнейшим нашим действием будет выбор альтернативной гипотезы, которая будет заключаться в том, что логарифмическая доходность рассматриваемых компаний распределена по закону Стьюдента. Проверим данное предположение на модельных данных. Используя ранее сгенерированную таблицу квантилей, находим вероятность ошибки второго рода, а после – мощность критерия, используя факт того, что в сумме они дают 1. Оценим мощность критерия при разных уровнях значимости, а именно 0,5%, 1%, 5%, 20%. Получаем следующие результаты

**2.  Проверка гипотезы для реальных данных**

**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.

Приложение 1

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

Время выполнения программ:

Приложение 2

Коды используемых программ