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

Департамент анализа данных,
принятия решений и финансовых технологий

Дисциплина: «Теория вероятностей и математическая статистика»
Направление подготовки: «Прикладная математика и информатика»
Профиль: «Анализ данных и принятие решений в экономике и финансах»
Факультет прикладной математики и информационных технологий
Форма обучения очная
Учебный 2019/2020 год, 4 семестр

Курсовая работа на тему:

«Проверка гипотезы о нормальном распределении логарифмической доходности при условии определенного уровня объема торгов накануне»

Вид исследуемых данных:
«Котировки акций компаний, входящих в индекс Dow Jones»

Выполнил:
студент группы ПМ18-5
Семенов А. М.
Научный руководитель:
Доцент, к.ф.-м.н.
Путко Б.А.

Москва 2020

Содержание

Введение

В данной курсовой работе будет проверена гипотеза о нормальном распределении логарифмической доходности при условии определенного уровня объема торгов накануне. Для анализа будем использовать акции индекса Dow Jones. Dow Jones - фондовых индексов, созданных Чарльзом Доу. Данный индекс включает компании из множества отраслей экономики. Мною были отобраны 10 компаний из разных отраслей экономики, которые входят в листинг BATS Global Market. Исследуемый период с 19 апреля 2010 года до 19 апреля 2020 года. Анализ проводится с использованием языка программирования Python на платформу Jupyter Notebook, что позволяет произвести анализ данных и графически изобразить полученные результаты анализа.

В первой части работы приводится теоретическая справка, где приводятся определения, которые касаются статистических гипотез и применяемых в данной работе критериев.

Во второй части рассматривается практическая часть данного вопроса, где применяются гипотезы для реальных и модельных данных.

В качестве новизны данной работы будет являться нестандартное, предложенное мной разбиение на уровни объемов торгов, к тому же для проверки логарифмической доходности на нормальность используются два критерия, которые будут сравнены между собой.

Я предполагаю, что логарифмическая доходность будет распределена не по нормальному закону, так как это заведомо верно.

1. Предварительный анализ данных

1.1. Тикеры компаний

Для исследования используются акции компаний, которые входят в индекс Dow Jones и торгуются на фондовой бирже BATS Global Market. Список компаний был взят с сайта «Банк Открытие» [1] и сайта группы «ФИНАМ» [2]. Компаний, вошедших в исследование, насчитывается десять штук, их список и соответствующие тикеры представлены нижу:

Таблица 1. Список компаний

Тикер	Компания
AAPL	Apple Inc.
AXP	American Express Co.
BA	Boeing Co.
CAT	Caterpillar Inc.
IBM	IBM Corp.
INTC	Intel Corp.
KO	Coca-Cola Co.
MSFT	Microsoft Corp.
V	Visa Inc.
XOM	Exxon Mobile Corp.

1.2. Количество торговых дней

По данным из предыдущего пункта, следует выяснить количество торговых дней для исследуемых компаний в рассматриваемом периоде. В нашем случае таких периодов несколько 18.04.2010 – 18.04.2020 (10 лет) и 18.04.2018 – 18.04.2020 (2 года). Выбраны два периода времени, чтобы проанализировать как будет отличаться гипотеза о нормальном распределении в них. Данная задача выполнена в программе «Рис.1,2,3,4,5.Предварительный анализ.ipynb». Результат представлен ниже:

Рисунок 1. Таблица числа торговых дней

	Тикер	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	AAPL	180	227	249	252	251	246	250	251	251	251	73
1	AXP	180	252	249	252	252	246	249	251	251	251	73
2	BA	180	246	249	252	252	246	249	251	251	251	73
3	CAT	180	246	248	252	252	246	249	251	251	251	73
4	IBM	180	246	249	252	252	246	249	251	251	251	73
5	INTC	180	246	249	252	251	246	249	251	251	251	73
6	KO	180	246	248	252	252	246	249	251	251	251	73
7	MSFT	180	246	249	252	251	246	249	251	251	251	73
8	V	0	25	249	148	235	251	58	42	251	251	73
9	XOM	180	246	248	252	252	246	249	251	251	251	73

Из результатов выполнения программы, которые представлены выше, можно увидеть, что не все компании оказались достаточно ликвидными. Из списка исследуемых данных следует исключить компанию с тикером V, так как компания с соответствующем тикером была ликвидной не на всем рассматриваем промежутке.

Далее необходимо рассмотреть таблицы максимальных отклоненных цен на акции, которые состоят из максимальных дневных скачков вверх и вниз по годам и тикерам, соответствующих компаний. Для каждой исследуемой компании берем из загруженных таблиц цену закрытия и считаем максимальные относительные скачки логарифмической доходности, затем у каждого тикера находим его максимальный относительный скачок в рассматриваемом периоде.

Рисунок 2. Таблица максимальные скачки вверх

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Макс скачок вверх
AAPL	0,074056	0,05013	0,084724	0,05042	0,07858	0,055874	0,063225	0,060382	0,06922	0,06617	0,11332	0,113320054
AXP	0,059755	0,066497	0,035339	0,050225	0,035806	0,060891	0,086739	0,057489	0,07341	0,044089	0,198465	0,198465448
BA	0,062175	0,051378	0,051512	0,052914	0,034526	0,059817	0,046618	0,094112	0,064377	0,061239	0,217817	0,217817161
CAT	0,071309	0,076306	0,045366	0,031886	0,05547	0,071527	0,074302	0,074993	0,064384	0,053625	0,102022	0,102021614
IBM	0,04462	0,053872	0,042585	0,048646	0,037004	0,044735	0,048771	0,085391	0,03725	0,083045	0,109447	0,10944671
INTC	0,056559	0,074683	0,032968	0,038175	0,090437	0,063437	0,034124	0,070442	0,100315	0,079674	0,175008	0,175007902
KO	0,025678	0,036848	0,022842	0,05559	0,037013	0,03934	0,024347	0,018372	0,029174	0,058935	0,061755	0,061755204
MSFT	0,051482	0,037342	0,055022	0,070926	0,038186	0,103713	0,055855	0,062266	0,074124	0,045675	0,133738	0,13373831
XOM	0,037169	0,049673	0,033347	0,031135	0,030449	0,044349	0,050187	0,021021	0,045764	0,03646	0,121614	0,121614391

Рисунок 3. Таблица максимальные скачки вниз

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Макс скачок вниз
AAPL	-0,04556	-0,08878	-0,06659	-0,13197	-1,93034	-0,05249	-0,06837	-0,0419	-0,06979	-0,10541	-0,14109	-1,930338515
AXP	-0,06815	-0,09246	-0,04404	-0,03675	-0,03745	-0,06645	-0,13098	-0,02291	-0,05794	-0,03785	-0,16029	-0,160292422
BA	-0,06539	-0,1064	-0,03631	-0,0481	-0,05319	-0,04409	-0,09377	-0,02891	-0,06615	-0,07054	-0,27215	-0,27215046
CAT	-0,0567	-0,11271	-0,05187	-0,06176	-0,05109	-0,07437	-0,06907	-0,04313	-0,07844	-0,09579	-0,15165	-0,151654758
IBM	-0,04006	-0,05374	-0,05004	-0,08736	-0,07455	-0,06121	-0,05821	-0,0509	-0,08011	-0,05724	-0,13836	-0,138362261
INTC	-0,04459	-0,06401	-0,0392	-0,06648	-0,0525	-0,04976	-0,09695	-0,03654	-0,09021	-0,09404	-0,19644	-0,19643769
KO	-0,0302	-0,03612	-0,69429	-0,03231	-0,06151	-0,0298	-0,04948	-0,02106	-0,04066	-0,08792	-0,10156	-0,69428982
MSFT	-0,04201	-0,08042	-0,04196	-0,12135	-0,04049	-0,09698	-0,07363	-0,03848	-0,05697	-0,03769	-0,16058	-0,160583601
XOM	-0,03731	-0,08629	-0,03919	-0,02941	-0,04307	-0,05376	-0,04254	-0,02016	-0,05755	-0,03965	-0,12966	-0,129656839

Из таблиц, которые получили в программе указанной выше можно подчеркнуть, что наибольшее относительное однодневное повышение цены замечено у компании с тикером BA в 2020 году. В свою очередь максимальное относительное однодневное понижение цены наблюдается у компании с тикеров AAPL в 2014 году. Ниже приведены графики цен на акции этих компаний за весь рассматриваемый период, которые были получены в указанной прогармме.

Рисунок 4. График цены акции компании с тикеров BA

Рисунок 5. График цены акции компании с тикером AAPL

2. Теоретическая справка по проверке гипотез

2.1. Статистическая проверка гипотез

Применение статистического критерия может привести к ошибкам двух различных типов:

1) Ошибка первого рода состоит в том, что отвергается верная гипотеза H_0 (основная);

2) Ошибка второго рода состоит в том, что отвергается верная гипотеза H_1 (альтернативная);

Вероятность ошибки первого рода называется *уровнем значимости* критерия и обозначается α. Вероятность ошибки второго рода обозначается β, а величина 1-β называется *мощностью критерия.* [3]

2.2. Критерий Колмогорова

Критерий Колмогорова часто используется на практике, так как он удобен в применении. За статистику критерия Колмогорова (d) принимают максимум по модулю разности значений эмпирической функции распределения $F_n(x)$ и теоретической функции распределения $F(x)$:

$$d = \max_{-\infty \le x \le \infty} |F_n(x) - F(x)|$$

Нулевая гипотеза H_0 принимается на уровне значимости α, если $\lambda = d\sqrt{n}$ удовлетворяет условию $\lambda \le \lambda_\alpha$.

Так как данный критерий удобен в применении, его часто используют на практике, при условии $n \ge 20$, так как при малых n фактический уровень значимости заметно отличается от номинального значения α.

2.3. P-значение критерия

P-значение часто используется в статистических программах, потому что оно позволяет решить вопрос о принятие или отвержении основной гипотезы одновременно для всех уровней значимости без вычисления критических значений, то есть она является альтернативой классической процедуре проверки.

Определение. Для фиксированной реализации $\vec{x}$ случайной выборки $\vec{X} = (X_1, \dots, X_n)$ P-значением статистического критерия называется такое число $\mathrm{PV}(\vec{x})$, что $\mathrm{PV}(\vec{x}) \ge \alpha$

для любого уровня значимости α, при котором гипотеза H_0 принимается, и $PV(\vec{x}) \leq \alpha$ - для любого уровня значимости α , при котором гипотеза H_0 отвергается. [3]

При верной гипотезе H_0 Р-значения равномерно распределены на отрезке [0;1].

2.4. Критерий Шапиро-Уилка

Критерий Шапиро-Уилка может применяться, когда рассматриваемая нами выборка имеет объем от 3 до 50, но в стандартном применении выборка должна состоять минимум из 8 величин. Однако с современными программными средствами данный критерий может применяться с выборками большего объема.

Основная гипотеза H_0 по данному критерию заключается в том, что исследуемая выборка распределена по нормальному закону. Альтернативная гипотеза H_1 – в том, что выборка не распределена по нормальному закону.

Пусть есть вариационный ряд $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, который построен по выборке $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$. Применяется формулы статистки данного критерия, которая имеет вид:

$$W = \frac{1}{S^2}\left[\sum_{i=1}^{n} a_{n-i+1} * (x_{n-i+1} - x_i)\right]^2,$$

где $S^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Числитель является квадратом оценки среднеквадратичного отклонения.

Коэффициенты a_{n-i+1} берутся из таблицы соответствующей таблицы, которую можно найти по ссылке, указанной в списке литературы.

Если W < W(α), то нулевая (основная) гипотеза о нормальном распределении отклоняется. Значение W(α) находится по таблице.

Критерий Шапиро-Уилка очень мощный для проверки нормальности, но он ограничен. При больших значения n таблицы коэффициентов a_{n-i+1} сложно применяются. [4]

2.5. ритерий нормальности Д’Агостино

Критерий Д’Агостино в основном применяют, когда отсутствуют сведения об альтернативном распределении.

Автор данного критерия предложил в качестве статистики для проверки нормальности распределения применить отношение оценки Даутона для стандартного отклонения к выборочному стандартному отклонению, оцененному методом максимального правдоподобия,

$$D = \frac{T}{sn^2},$$

где $T = \sum_{i=1}^{n} \left\{ i - \frac{(n+1)}{2} \right\} x_i, x_1 \leq \cdots \leq x_n; s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$

Следует отметить, что $\bar{\sigma} = \frac{2\sqrt{\pi}T}{n(n-1)}$ является несмещенной оценкой стандартного отклонения σ.

В качестве статистики критерия Д'Агостино принимают следующую величину

$$Y = \sqrt{n} \frac{D - 0{,}28209473}{0{,}02998598}.$$

Гипотеза нормальности принимается в том случае, если $Y_1(\alpha) \leq Y \leq Y_2(\alpha)$, где $Y_1(\alpha)$ и $Y_2(\alpha)$ – критическое значение статистики Y при достоверности α.

Данный критерий уступает лишь немного уступает по мощности критерию Шапиро-Уилка. [5]

3. Практическая часть

3.1. Проверка гипотезы для модельных данных

В программе «Таб.2;Рис.6.Модельные данные.ipynb» строится нормальное распределение для выборки объема 252 и на основе его высчитываются квантили уровней 0,1; …; 0,9, которые сохраняются в файл «Таблица_квантилей_из_выборки_объёма 252.xlsx». Кроме этого строятся квантили уровня 0,111; …; 0,999.

Таблица 2. Квантили

Уровень	Квантиль
0,1	-0,36922
0,2	-0,27086
0,3	-0,19452
0,4	-0,12697
0,5	-0,05885
0,6	0,007759
0,7	0,087251
0,8	0,19241
0,9	0,355151

Далее осуществляется проверка на равномерность распределения Р-значения на отрезке от 0 да 1. Данная проверка производится в программе, указанной выше. Сначала 10000 раз строится нормальное распределение, от которого находим Р-значение по основному критерия в данной работе критерию Шапиро-Уилка.

Рисунок 6. Р-значения для модельных данных

Из рисунка выше ожидаемо следует, что гипотеза для модельных данных о нормальном распределении принимается, так как равномерность подтверждена графически.

3.2. Выбор альтернативной гипотезы для оценки мощности критерия

В качестве альтернативных гипотез мною были выбраны распределения Стьюдента со степенями свободы 2, 4 и 6. Данное распределение широко применятся в задачах, где требуется обработка экспериментальных данных. Кроме этого распределение Стьюдента при увеличении степеней свободы начинает совпадать с нормальным [5].

В программе «Таб.3,4,5.Мощность критерия.ipynb» 1000 раз строится распределение Стьюдента по очереди для разных степеней свободы объема равному кварталу, полугодию и году. Далее на его основе вычисляется мощность критерия. Мощность равна отношению количества P-значений меньших 0,05 к общему числу построений рассматриваемого распределения.

Таблица 3. Мощность критерия для распределения Стьюдента (n=2)

	Квартал	**Полугодие**	**Год**
Мощность	0,936	0,996	1

Таблица 4. Мощность критерия для распределения Стьюдента (n=4)

	Квартал	**Полугодие**	**Год**
Мощность	0,532	0,783	0,973

Таблица 5. Мощность критерия для распределения Стьюдента (n=6)

	Квартал	**Полугодие**	**Год**
Мощность	0,356	0,523	0,77

Из получившихся выше таблиц следует сделать вывод, что мощность критерия возрастает при увлечении интервалов времени. Следует добавить, что при небольших степенях свободы полученная мощность достаточно высокая, отсюда следует, что критерий Шапиро-Уилка является мощным и вероятность ошибки второго роды низкая. Однако при увеличении степеней свободы мощность заметно уменьшается, что является неблагоприятным знаком для критерия Шапиро-Уилка, так как его мощность падает и увеличивается вероятность ошибки второго рода.

3.3. Проверка гипотезы для реальных данных

Сначала при работе с реальными данными будет произведена проверка для логарифмической доходности без объема торгов. В программе «Рис.7,8,9,10,11,12.Рельные

данные.ipynb» высчитываются логарифмические доходности компаний за определенный год и для полученных данных вычисляется Р-значение критерия Шапиро-Уилка и для сравнения критериев к тому же вычисляется критерий Д’Агостино.

Из получившихся значений критериев вычисляются медианные значения по компаниям за весь период и по всем компаниям за определенный год.

Рисунок 7. Таблица Р-значений критерия Шапира-Уилка

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Медиана
AAPL	0,00066	0,00012	0	0	0	0,07212	0	0	0,00003	0	0,00536	0
AXP	0,19641	0	0,58363	0,00733	0,00343	0	0	0	0	0,0001	0,00018	0,0001
BA	0,19442	0	0,00002	0,03735	0,00012	0	0	0	0,04371	0,00011	0,00002	0,00002
CAT	0,01175	0,00157	0,24771	0	0	0	0,00015	0	0,00078	0	0,03668	0,00015
IBM	0,00294	0,00006	0	0	0	0	0	0	0	0	0,09475	0
INTC	0,12157	0,01139	0,09702	0	0	0,00468	0	0	0,00009	0	0,00047	0,00009
KO	0,00007	0,01398	0	0	0	0,00533	0	0,00004	0,00013	0	0,00774	0,00004
MSFT	0,01176	0,00002	0,00071	0	0,00254	0	0	0	0	0,00103	0,00123	0,00002
XOM	0,01465	0	0,00026	0,07665	0,00001	0,00005	0	0,01657	0,00018	0,18932	0,01221	0,00026
Медиана	0,01176	0,00006	0,00026	0	0	0	0	0	0,00009	0	0,00536	---

Рисунок 8. Таблица Р-значений критерия Д’Агостино

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Медиана
AAPL	0,0003	0	0	0	0	0,08629	0	0	0,00059	0	0,02538	0
AXP	0,27261	0	0,12813	0,00256	0,04429	0	0	0	0	0,00048	0,0024	0,00048
BA	0,11605	0	0,00001	0,01301	0	0	0	0	0,04583	0,00024	0,00965	0,00001
CAT	0,04877	0,00002	0,19487	0	0	0	0,00016	0	0,00005	0	0,0048	0,00002
IBM	0,00749	0,00013	0	0	0	0	0	0	0	0	0,05586	0
INTC	0,06004	0,0021	0,22451	0	0	0,00051	0	0	0,00003	0	0,00175	0,00003
KO	0,01139	0,02508	0	0	0	0,00067	0	0,00001	0	0	0,04849	0
MSFT	0,02777	0	0,00004	0	0,01508	0	0	0	0,00009	0,00415	0,00318	0,00004
XOM	0,0691	0	0,00007	0,03168	0	0,0017	0,00001	0,26719	0	0,10373	0,08144	0,0017
Медиана	0,04877	0	0,00004	0	0	0	0	0	0,00003	0	0,00965	---

По результатам программы следует заметить, что экстремальные медианные значения наблюдаются в обоих критериях по годам 2010 и 2020, а также в критерии Шапиро-Уилка по компаниям с тикерами CAT и XOM, а в Д’Агостино по AXP и XOM. На мой взгляд значения двух критериев отличаются прежде всего тем, что критерий Шапиро-Уилка более мощный, значит при его использовании вероятность ошибки второго рода будет ниже.

Далее в данной программе рассматривается распределение Р-значений.

Рисунок 9. Гистограмма P-значений для реальных данный (Шапиро-Уилка)

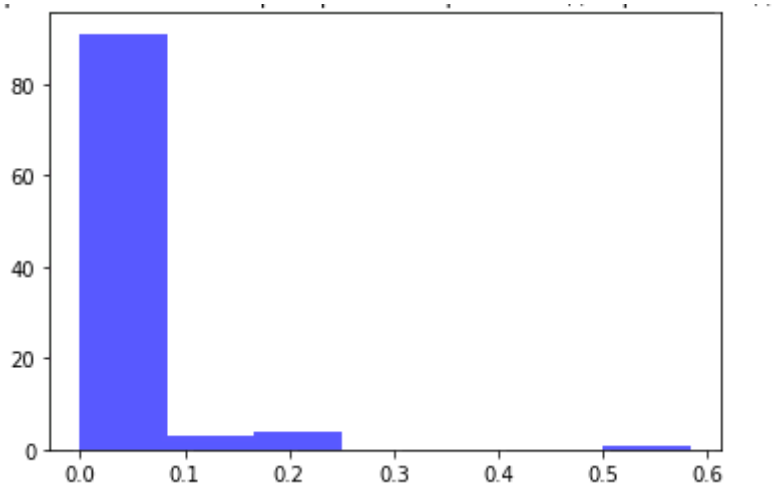
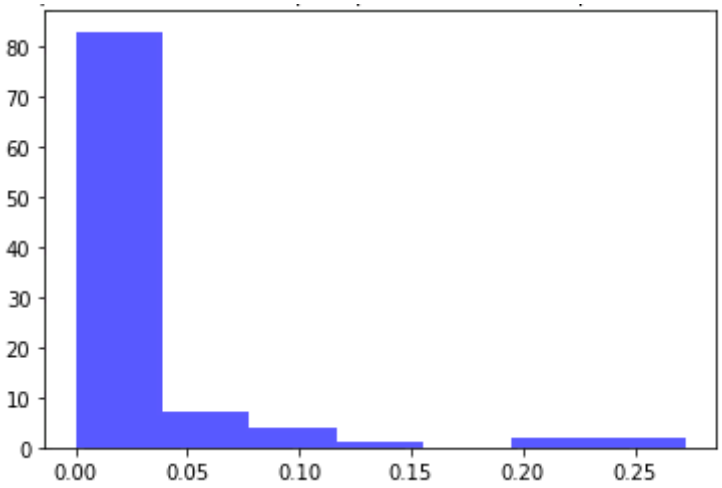


Рисунок 10. Гистограмма P-значений для реальных данный (Д'Агостино)



При рассмотрении получившихся гистограмм сразу следует вывод, что P-значения в обоих критериях распределены неравномерно. Таким образом, гипотеза о нормальности распределения логарифмической доходности отвергается.

В программе из данного пункта к тому же считается доля принятий гипотезы о нормальном распределении при уровнях значимости 0,05 и 0,01:

Рисунок 11. Доля выполнения гипотезы (Шапиро-Уилка)

	уровень значимости 5%	уровень значимости 1%
Доля	0.10101	0.20202

Рисунок 11. Доля выполнения гипотезы (Д'Агостино)

	уровень значимости 5%	уровень значимости 1%
Доля	0.121212	0.232323

Доля принятий гипотезы по обоим критериям при разных уровнях значимости мала и примерно располагается на одном уровне.

3.3.1. Проверка гипотезы о нормальном распределении дневной логарифмической доходности при условии определенного объема торгов накануне

Сначала объем торгов разделяется на малый, средний и большой. Основная гипотеза будет проверяться по критерию Шапиро-Уилка в программе «Таб.6,7.Уровень объема торгов.ipynb». Объем торгов вычисляется из произведения полу-суммы цены открытия и закрытия и объема торгов в штучном выражении. После чего высчитывается логарифмическая доходность из получившихся цен. Новшеством в моей работы является, что объем торгов был разделен не стандартным образом: относительно квартала считается логарифмическая доходность по всем кварталам за рассматриваемый период, из получившихся доходностей находится среднее значение логарифмической доходности для каждого квартала и от средних квартальных значений вычисляется Р-значение по критерию Шапиро-Уилка; для двухгодичного периода Р-значения критерия выселяются аналогично; для всего периода берется логарифмическая доходность акции за весь период от нее находится Р-значение. Далее происходит проверка на принятие или отвержение основной гипотезы.

Я предполагал, что Р-значения получатся примерно одинаковыми для разных объемов торгов, так как для меньших объемов берутся усредненные значения. Однако получились со всем иные результаты, которые можно увидеть ниже.

Таблица 6. Р-значения при определенном уровне объема торгов

	Малый	Средний	Большой
AAPL	0,00035	0,05018	0
AXP	0,23434	0,96921	0

BA	0,017	0,27855	0
CAT	0,00117	0,90719	0
IBM	0,02118	0,59729	0
INTC	0,98631	0,69809	0
KO	0,00013	0,2193	0
MSFT	0,28608	0,96516	0
XOM	0,00024	0,21443	0

Таблица 7. Принятие гипотезы при определенном уровне объема торгов

	Малый	**Средний**	**Большой**
AAPL	-	+	-
AXP	+	+	-
BA	-	+	-
CAT	-	+	-
IBM	-	+	-
INTC	+	+	-
KO	-	+	-
MSFT	+	+	-
XOM	-	+	-

Таким образом, можно заметить, что в большинстве случаев гипотеза отвергается. Однако следует интересный результат, что при данном разбиении средний уровень принимается весь. Значит, при предложенном мной разбиении после анализа результатов напрашивается вывод, что нормальность распределения логарифмической доходности вероятнее всего зависит от уровня объема торгов накануне.

Заключение

В данной курсовой работе с помощью языка программирования Python была проведена проверка гипотезы о нормальном распределении логарифмической доходности акций компаний, входящих в листинг индекса Dow Jones, при условии определенного объема торгов накануне. Проверка была осуществлена с использованием критерия Шапиро-Уилка и критерия Д'Агостино. Новизна данной работы заключалась в принципе изменения разбиения объемов выборок и написании программ на языке Python.

В данной работе применялись программные средства, которые напрямую считаю значения статистик и P-value в отличии от работ прошлых лет, где вычисления велись через программный код.

По результатам проведенной работы следует сделать вывод, что гипотеза о нормальности распределения логарифмической доходности акций при условии определённого уровня торгов накануне в большинстве случаев отвергается, вне зависимости от объема торгов при предложенном мной разбиении. Результат подтверждает некоторые итоги курсовых работ прошлых лет.

Список используемых источников

1. https://journal.open-broker.ru/economy/chto-takoe-index-dow-jones/

2. https://www.finam.ru

3. *Браилов А.В.* Лекции по математической статистике. М.: Финакадемия, 2007, 172с.

4. http://www.machinelearning.ru/wiki/index.php?title=Критерий_Шапиро-Уилка

5. *Кобзарь А.И.* Прикладная математическая статистика. Для инжинеров и научных работников. М.: ФИЗМАТЛИТ, 2006, 816 с.

6. Курсовые работы прошлых лет.

Приложения

Приложение 1.

Характеристики компьютера:

- Тип процессора: Intel Pentium CRU B960
- Тактовая частота: 2.2GHz
- Частота системной шины: 100МГц
- Объем кэша второго уровня 256 КБ per core (On-Die, ECC, Full-Speed)

Список программ, работающих более 10 секунд:

- Рис.1,2,3,4,5.Предварительный анализ.ipynb
- Таб.2;Рис.6.Модельные данные.ipynb
- Рис.7,8,9,10,11,12.Рельные данные.ipynb
- Таб.6,7.Уровень объема торгов.ipynb

Приложение 2.

- **Рис.1,2,3,4,5.Предварительный анализ.ipynb**

```
#Семенов 2020 год
#Примерное время выполнения 17 секунд
import pandas as pd #библиотека для формирования таблицы
import time #библиотке для времени работы программы
start = time.time()
def ndays(year, file): #функция для вывода числа торговых дней за
один год
    df = pd.read_csv(file, sep=';', encoding='cp1251') #считываем
csv файл
    condition = (df['<DATE>']>=year*10000) &
(df['<DATE>']<(year+1)*10000) #условие для одного года
    return len(df[condition])
tickers = ['AAPL', 'AXP', 'BA', 'CAT', 'IBM', 'INTC', 'KO', 'MSFT',
'V', 'XOM']
years = range(2010,2021)
dfNDays = pd.DataFrame() #создаём объект таблтцу
dfNDays['Тикер'] = tickers
for year in years: #проход по годам
    yearDays = [] #список кол-ва дней для всех компаний за один год
    for ticker in tickers:
        f = open(u'C:/Users/Андрюша/Desktop/Акции_курсовая/' +
ticker + '.csv') #открываем файл
        yearDays.append(ndays(year, f))
    dfNDays[str(year)] = yearDays #заполняем таблицу
dfNDays.to_excel(u'C:/Users/Андрюша/Desktop/Акции_Курсовая/Число_тор
говых_дней.xlsx', index = False,  encoding = 'cp1251')
print("Примерное время выполнения = {}".format(time.time() - start))
dfNDays
import numpy as np
import scipy.stats as st
def ndays(year, file): #функция для вывода логдоходностей
```

```
    df1 = pd.read_csv(file, sep=';', encoding='cp1251') #считываем
csv файл
    condition = (df1['<DATE>']>= year*10000) & (df1['<DATE>'] <
(year+1)*10000) #условие для одного года
    df = df1[condition]['<CLOSE>'] #отбираются цены закрытия для
данного года
    df = pd.DataFrame.diff(np.log(df1[condition]['<CLOSE>']))[1::]
#подсчет логдохоности
    return [max(df), min(df)] #Возвращает список, состоящий из мак-
симальных скачков вверх и вниз логдоходностей для компании за данный
год
tickers = ['AAPL', 'AXP', 'BA', 'CAT', 'IBM', 'INTC', 'KO', 'MSFT',
'XOM']
dfNDays_max = pd.DataFrame(index = tickers) #создаём объект таблтцу
max скачки вверх логдоходности, где индексы это года
dfNDays_min = pd.DataFrame(index = tickers) #создаём объект таблтцу
max скачки вверх логдоходности, где индексы это года
for year in range(2010, 2021):
    p = [] #список занчений мин и макс логдоходности для всех компа-
ний и одного года
    for ticker in tickers:
        f = open(u'C:/Users/Андрюша/Desktop/Акции_курсовая/' +
ticker + '.csv') #открываем файл
        p.append(ndays(year, file = f)) #добавляем в список
    dfNDays_max[str(year)] = [i[0] for i in p] #отбираем из создан-
ного списка max скачки вверх логдоходности и заполяем таблицу
    dfNDays_min[str(year)] = [i[1] for i in p] #отбираем из создан-
ного списка max скачки вниз логдоходности и заполяем таблицу
maxi = [] #max скачки вверх логдоходности за весь период для всех
компаний
mini = [] #max скачки вниз логдоходности за весь период для всех
компаний
for i in range(0, len(tickers)): #проходим по всем компаниям у каж-
дой max вверх и вниз скасчки логдоходностей
    maxi.append(max(dfNDays_max.iloc[i]))
    mini.append(min(dfNDays_min.iloc[i]))
dfNDays_max["Макс скачок вверх"] = maxi
dfNDays_min["Макс скачок вниз"] = mini
dfNDays_max.to_excel(u'C:/Users/Андрюша/Desktop/Акции_Курсовая/Макс_
скачок.xlsx', index = True,  encoding = 'cp1251')
dfNDays_min.to_excel(u'C:/Users/Андрюша/Desktop/Акции_Курсовая/Мин_с
качок.xlsx', index = True,  encoding = 'cp1251')
for i in range(0, len(tickers)): #цикл для определения max из всех
макс скачков вверх и min из всех max вниз
    if max(dfNDays_max["Макс скачок вверх"]) ==
dfNDays_max.iloc[i]["Макс скачок вверх"]:
        max_tick = tickers[i]
```

```
        print('Наибольшее относительное однодневное повышение цены =
{}, наблюдается у компании с тикером
{}'.format(max(dfNDays_max["Макс скачок вверх"]), max_tick))
import matplotlib.pyplot as plt
from numpy import linspace
tickers = ["BA", "AAPl"] #такеры, график которых необходимы
def Close(y1, y2, file):
    csvtab = pd.read_csv(file, sep=';', encoding='cp1251')
    df = pd.DataFrame()
    df['date'] = csvtab['<DATE>']
    df['close']= csvtab['<CLOSE>']
    condition = (df['date']>=y1*10000) & (df['date']<(y2+1)*10000)
    return df['close'][condition]
for ticker in tickers:
    f = open(u'C:/Users/Андрюша/Desktop/Акции_курсовая/' + ticker +
'.csv') #открываем файл
    y = Close(2010, 2020, f)
    x = linspace(2010, 2020, len(y))
    fig, ax = plt.subplots(figsize=(10, 5))
    plt.title('График цены закрытия компании с тикером ' + ticker +
' за весь исследуемый период', size = 20)
    plt.grid(True)
    plt.plot(x, y)
    plt.savefig(u"C:/Users/Андрюша/Desktop/Акции_Курсовая/График це-
ны закрытия компании с тикером " + ticker + " за весь исследуемый
период")
```

- **Таб.2;Рис.6.Модельные данные.ipynb**

```
#Курсовые работы прошлых лет, Семенов 2020 год
#Примерное время выполнения 200 секунд
import pandas as pd
import numpy as np
import scipy.stats as st
from scipy.stats import mstats as mst
import math
import matplotlib.pyplot as plt
import time
n = 252 #объём выборки
m = 10000 #количество произведенных вычислений
ex= [0] * m
mod_set = [[]] * m #список нормальных распределений на каждом шаге
for i in range(m):
    mod_set[i] = np.random.normal(size=n)
    ex[i] = st.kurtosis(mod_set[i], fisher=True, bias=True)
kvant = [0] * 9 #для квантилей от 0,1 до 0,9
for i in range(1,10):
    kvant[i-1] = float(mst.mquantiles(ex, prob=[i/10], betap=0.5,
alphap=0.5))
a = range(1, 1000, 1) #для квантилей от 0,001 до 0,999
```

```
qs = [0] * len(a)
for i in a:
    qs[i-1] = float(mst.mquantiles(ex, prob=[i/1000], betap=0.5,
alphap=0.5))
tab_qvant = pd.DataFrame() #создаем таблицу
tab_qvant['Уровень'] = [i/10 for i in range(1,10)]
tab_qvant['Квантиль'] = kvant
tab_qvant.to_excel(u'C:/Users/Андрюша/Desktop/Акции_курсовая/Таблица
_квантилей_из_выборки_объёма ' + str(n) + ".xlsx", index=False, en-
coding='cp1251')
pv=[] # список p-value для гистограммы
for i in range(m):
    pv.append(st.shapiro(np.random.normal(size=n))[1])
kolint=(math.log2(len(pv)))//1+1
plt.title('Гистограмма P-значений для модельных данных')
plt.hist(pv, bins=int(kolint), color='b', alpha=0.65)
plt.savefig(u"C:/Users/Андрюша/Desktop/Акции_Курсовая/Гистограмма P-
значений для модельных данных.png")
print("Стандартная ошибка {}".format(st.sem(ex)))
```

- **Таб.3,4,5.Мощность критерия.ipynb**

```
#Семенов 2020 год
import pandas as pd
import numpy as np
import scipy.stats as st
def mosh(n = 2): #Функция возвращает таблицу распределения Стъюдента
    m = 1000
    per = [63, 126, 252] #периоды: квартал, полугодие, год
    k = ['Квартал', 'Полугодие', 'Год'] #назвние полей в таблице
    t = pd.DataFrame(index = ['Мощность']) #таблица
    for i in per:
        val_k = 0 #счётчк значений p-value < 0.05
        for j in range(m):
            if st.shapiro(np.random.standard_t(n, size=i))[1] <
0.05:
                val_k += 1
        t[k[per.index(i)]] = val_k / m # [заполняем соответсвующие
поля таблицы]

t.to_excel(u'C:/Users/Андрюша/Desktop/Акции_курсовая/Мощность_критер
ия_для_распределения_Стъюдента(n = ' + str(n) + ').xlsx', in-
dex=True, encoding='cp1251')
    return t
mosh() #распредление Стюдентна со степенями свободы = 2
```

- **Рис.7,8,9,10,11,12.Рельные данные.ipynb**

```
#Семенов 2020 год
#Примерное время выполнения 11 секунд
import math
```

```
def log_dog(year, file): #функция для вывода логдоходностей компании за один год
    df1 = pd.read_csv(file, sep=';', encoding='cp1251') #считываем csv файл
    condition = (df1['<DATE>']>= year*10000) & (df1['<DATE>'] < (year+1)*10000) #условие для одного года
    df = df1[condition]['<CLOSE>'] #отбираются цены закрытия для данного года
    df = pd.DataFrame.diff(np.log(df1[condition]['<CLOSE>']))[1::] #подсчет логдохоности
    return df #Возвращает таблицу лог_дохоностей
def p_value(lg): #Функция для расчета p-value
    p_sp = list()
    p_sp.append(round(st.shapiro(lg)[1], 5)) #Шапиро-Уилка
    p_sp.append(round(st.normaltest(lg)[1], 5)) #Агадино
    return p_sp
def dol(p_val): #количество значений, которые принимаются с разными уровнями значимости 0.01 и 0.05
    k_05 = 0
    k_01 = 0
    for i in p_val:
        if i >= 0.05:
            k_05 += 1
            k_01 += 1
        elif i >= 0.01:
            k_01 += 1
    return [k_05, k_01]
def pv_hist(pv, text = 'Шапиро-Уилка'): #Функция для построения Гит-сограммы p-value
    kolint=(math.log2(len(pv)))//1+1
    plt.title('Гистограмма P-значений критерия ' + text + ' для реальных данных')
    plt.hist(pv, bins=int(kolint), color='b', alpha=0.65)
    m = list() #список медиан для всех компаний
    for i in range(n):
        m.append(np.median(tabl.iloc[i]))
    m.append("---") #заполняем заключительную строку, так как медиана от медиан по годам НЕ НУЖНА!
    return m
tickers = ['AAPL', 'AXP', 'BA', 'CAT', 'IBM', 'INTC', 'KO', 'MSFT', 'XOM']
table_k = pd.DataFrame(index = ind)
k_sh = [[], []] #список количества принимаемых p-value для критерия Шапиро-Уилка по годам для разных уровней значимости
k_k = [[], []] #список количества принимаемых p-value для критерия Агадино по годам для разных уровней значимости
p_vse = [[], []] #Список p-value для каждого критерия по всем тикерам за весь период
```

```
for year in range(2010, 2021):
    p = [[], []] #список значений p-value
    for ticker in tickers:
        f = open(u'C:/Users/Андрюша/Desktop/Акции_курсовая/' + tick-
er + '.csv') #открываем файл
        l_g = log_dog(year, file = f) #добавляем в список

        p[0].append(p_value(l_g)[0]) #значение p-value для Шапиро-
Уилка
        p[1].append(p_value(l_g)[1]) #значение p-value для Агадино
        p_vse[0].append(p_value(l_g)[0])
        p_vse[1].append(p_value(l_g)[1])
    k_sh[0].append(dol(p[0])[0])
    k_k[0].append(dol(p[1])[0])
    p[0].append(np.median(p[0])) #Вычисляем медиану
    table_sh[str(year)] = p[0]
table_sh['Медиана'] = med(table_sh, len(tickers))
table_sh.to_excel(u'C:/Users/Андрюша/Desktop/Акции_Курсовая/P-
значения_Шапиро_Улка.xlsx', index = True,  encoding = 'cp1251')
table_sh #Таблица Шапиро-Уилка
dly_SHAPIR = pd.DataFrame(index = ['Доля']) #Таблица долей принятия
гипотезы H0 для Шапиро-Уилка
dly_SHAPIR['уровень значимости 5%'] = sum(k_sh[0])/(len(range(2010,
2021))*len(tickers))
dly_SHAPIR['уровень значимости 1%'] = sum(k_sh[1])/(len(range(2010,
2021))*len(tickers))
pv_hist(p_vse[0]) #Строим гистограмму для критерия Шапиро-Уилка
```

- **Таб.6,7.Уровень объема торгов.ipynb**

```
#Семенов 2020 год
#Примерное время выполнения 12 секунд
def p_v_kv(file): #функция для вывода p-value срзнач логдоходностей
по кварталам для одной компании
    df1 = pd.read_csv(file, sep=';', encoding='cp1251') #считываем
csv файл
    kv = [(1, 3), (4, 6), (7, 9), (10, 12)] #список месяцев начала и
конца кварталов
    df_mean = [] #среднее значение лог доходностей за каждый квартал
    for year in range(2011, 2019): #так как 2010 и 2020 года взяты
не полностью
        date = year*100 #формируем дату для условия одного квартала
        for i in kv:
            condition = (df1['<DATE>'] >= ((date + i[0])*100 + 1)) &
(df1['<DATE>'] <= ((date + i[1])*100 + 31)) #условие для одного
квартала
            df = pd.DataFrame()
            df['close'] = df1[condition]['<CLOSE>'] #отбираются цены
закрытия для данного периода
```

```
            df['open'] = df1[condition]['<OPEN>'] #отбираются цены
открытия для данного периода
            df['vol'] = df1[condition]['<VOL>'] #отбираются объем
торгов в шт для данного периода
            df['rvol'] = ((df['close'] + df['open']) / 2) *
df['vol']
            df = pd.DataFrame.diff(np.log(df['rvol'] ))[1::] #под-
счет логдохоности
            df_mean.append(df.mean()) #добавление срднего значения
за квартал
    return  round(st.shapiro(df_mean)[1], 5) #p-value по средним
квартальным логдоходностям для одной компании
def p_v_2_years(file): #функция для вывода p-value срзнач логдоход-
ностей для двух лет для одной компании
    …
def p_v_all(file): #функция для вывода p-value логдоходностей компа-
нии за весь период для одной компании
    …
def check(p_t): #формирование таблицы, которая показывает принимает-
ся гипотеза или нет
    ch = []
    for p in p_t:
        if p >= 0.05:
            ch.append('+')
        else:
            ch.append('-')
    return ch
tickers = ['AAPL', 'AXP', 'BA', 'CAT', 'IBM', 'INTC', 'KO', 'MSFT',
'XOM']
i = 0 #счётчик для индексов
#списки p-value для малого, среднего и большого объема торгов
p_mal = list()
p_sr = list()
p_big = list()
#списки принятия основной гипотезы для малого, среднего и большого
объема торгов
ch_mal = list()
ch_sr = list()
ch_big = list()
for ticker in tickers:
       with open(u'C:/Users/Андрюша/Desktop/Акции_курсовая/' + tick-
er + '.csv') as f:
            p_mal.append(p_v_kv(f))
       …
        ch = check([p_mal[i], p_sr[i], p_big[i]])
        i += 1
        ch_mal.append(ch[0])
        ch_sr.append(ch[1])
```

```
        ch_big.append(ch[2])
print("Примерное время выполнения = {}".format(time.time() - start))
P_znach_SHAPIR = pd.DataFrame(index = tickers)
P_znach_SHAPIR['Малый'] = p_mal #квартал
P_znach_SHAPIR['Средний'] = p_sr #2 года
P_znach_SHAPIR['Большой'] = p_big #2011-2019
P_znach_SHAPIR.to_excel(u'C:/Users/Андрюша/Desktop/Акции_Курсовая/P-значения_при_опр._уровне_объема_торгов.xlsx', index = True,  encoding = 'cp1251')
P_znach_SHAPIR
check_gyp = pd.DataFrame(index = tickers)
check_gyp['Малый'] = ch_mal #квартал
check_gyp['Средний'] = ch_sr #2 года
check_gyp['Большой'] = ch_big #2011-2019
check_gyp.to_excel(u'C:/Users/Андрюша/Desktop/Акции_Курсовая/Проверка_гипотезы_при_опр._уровне_объема_торгов.xlsx', index = True,  encoding = 'cp1251')
check_gyp
```

Приложение 3.

Файлы:

- Курсовая работа Семенов Андрей ПМ18-5.docx
- Курсовая_работа_Семенов_Андрей_ПМ18-5.pdf
- Отчет по антиплагиату 03_05_2020 23_20_01.pdf
- Рис.1,2,3,4,5.Предварительный анализ.ipynb
- Таб.2;Рис.6.Модельные данные.ipynb
- Таб.3,4,5.Мощность критерия.ipynb
- Рис.7,8,9,10,11,12.Рельные данные.ipynb
- Таб.6,7.Уровень объема торгов.ipynb
- Файлы тикеров 10 шт (.csv)
- Гистограмма P-значений Агостино для реальных данных
- Гистограмма P-значений для модельных данных
- Гистограмма P-значений Шапиро-Уилка для реальных данных
- График цены закрытия компании с тикером AAPl за весь исследуемый период
- График цены закрытия компании с тикером BA за весь исследуемый период
- P-значения_Агостино
- P-значения_при_опр._уровне_объема_торгов
- P-значения_Шапиро_Улка
- Макс_скачок
- Мин_скачок
- Мощность_критерия_для_распределения_Стъюдента(n = 2, 4, 6)
- Проверка_гипотезы_при_опр._уровне_объема_торгов
- Таблица_квантилей_из_выборки_объёма 252
- Число_торговых_дней