Федеральное государственное образовательное бюджетное учреждение
высшего профессионального образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)
Факультет «Факультет информационных технологий и анализа больших данных»
Кафедра «Теория вероятностей и математическая статистика»

Курсовая работа

на тему:

«Проверка гипотезы о нормальном распределении логарифмической доходности по критерию Шапиро - Уилка»

Вид исследуемых данных:

Котировки акций, входящих в индекс «DAX»

Выполнила:
студентка группы ПМ19-4
Качуляк Маргарита
Григорьевна

Научный руководитель:
профессор ДАДиМО, д.э.н
Коровин Дмитрий Игоревич

Москва 2021г.

Оглавление

1. Введение

В представленной курсовой работе будет проведена проверка гипотезы о нормальном распределении логарифмических доходностей, рассчитанных по котировкам акций, которые входят в индекс «DAX» (Deutsche Akzien Index) - важнейший фондовый индекс, который является одним из самых значимых индикаторов состояния фондового рынка Германии и всего Евросоюза. Основным критерием стал один из наиболее эффективных - критерий нормальности Шапиро – Уилка, заключенный в оценке линейной комбинации разностей порядковых статистик. Отобрано 10 тикеров. В качестве основного таймфрейма был использован период с 2016 по 2020 включительно с периодичностью в 1 день, однако для более точного прогнозирования также были проведены дополнительные исследования, в которых удалось рассмотреть временные интервалы до - в пик - после начала пандемии коронавируса и сделать некоторые выводы.

Гипотезы о нормальном распределении логарифмических доходностей являются первоначалом многих теорий фондовых рынков, в то время как этот факт сам по себе вызывает сомнения. Цель моего исследования – выяснить насколько гипотеза о нормальном распределении дневных логарифмических доходностей соответствует реальности, ведь выявление, пусть приближенного, соответствия распределения реальных значений какому-либо теоретическому распределению позволяет найти вероятность, с которой возникнет то или иное событие в обозримом будущем, откуда вытекает актуальность работы. Основная задача курсовой – создание универсального аппарата для применения его в торговой стратегии, а новизна заключаются в проверке наших данных на нормальность распределения логарифмической доходности в разные периоды пандемии. Отметим, что ранее не возникало таких экономических ситуаций, в ходе которых рынок менялся так сильно и стремительно.

Проверка критерия пройдет в несколько этапов. Сначала будет рассмотрена теория по выбранной теме. Затем необходимо провести предварительный анализ данных для исключения тикеров не подлежащих исследованию. Далее начнется практическая часть, в которой сперва гипотеза проверится на модельных данных, потом с помощью альтернативных гипотез будет найдена оценка мощности критерия. В последствие начнется работа с реальными данными.

2. Теоретическая справка

2.1. Логарифмическая прибыль и нормальное распределение с точки зрения экономической теории

Для расчёта доходности и показателя эффективности инвестирования с учетом риска на практике используются 2 подхода:

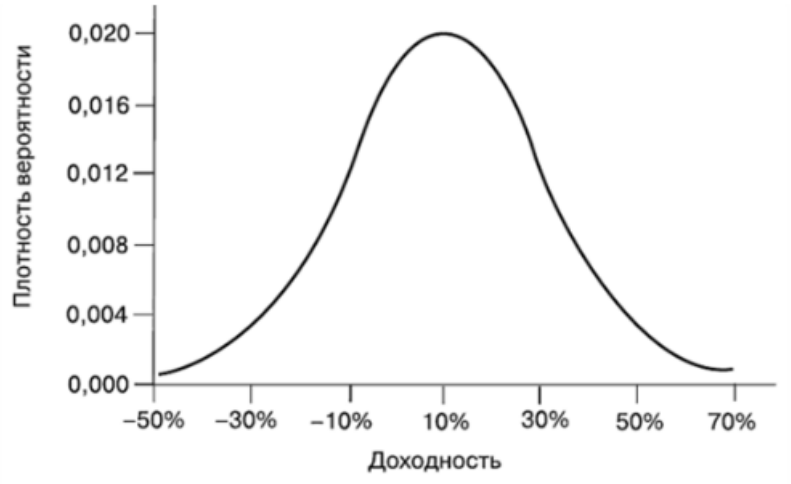
1. Процентная доходность с момента времени t до момента i: $r_{i,t}^{\%} = \frac{p_i - p_{i-t}}{p_{i-t}} = \frac{p_i}{p_{i-t}} - 1$, где p_i – цена в i-ый период времени.
2. Логарифмическая доходность с момента времени t до момента i: $r_{i,t} = \ln\left(\frac{p_i}{p_{i-t}}\right)$

Для работы с финансовыми активами чаще используется **логдоходность** – натуральный логарифм отношения конечного капитала к начальному капиталу (отношение конечной стоимости к начальной). Происходит это потому, что при подсчете приращения цены в процентах, если цена снижается на n%, а затем повышается на те же n%, то получается значение, не равное первоначальной цене, когда при применяя логдоходность если взять первый член ряда Маклорена от функции ln(P2/P1), будем иметь P2/P1-1 – равно как и обычная доходность (даже при малых изменениях).

Ценовой логарифм дает возможность уйти от искажений, которые связаны с абсолютными значениями цены, к относительной оценке. Это играет большую роль в инвестировании и трейдинге, так как для инвестора имеет вес не цена акции, а ее относительное изменение за какой-либо временной интервал. Поэтому вкладчика прежде всего беспокоит именно логарифмы цен, ведь они лучше отражают его восприятие прибыли и убытков.

Предположение о нормальности распределения доходности финансовых активов является одним из самых значимых показателей, дающих инвесторам оценивать риски при вложении средств в портфели и финансовым экономистам рассматривать этот косвенный показатель для предсказания доходности.

Рисунок отображает приблизительное поведение доходностей при выполнении условия нормальности распределения.



2.2. Статистическая гипотеза

Статистической принято называть гипотезу о виде неизвестного распределения генеральной совокупности или о параметрах известных распределений Ее проверку необходимо осуществляется путем применения методов математической статистики.

На практике чаще всего требуется проверить некоторую конкретную гипотезу **H0**, называемую **нулевой** (основной), и параллельно рассматривать противоречащую ей – альтернативную (**конкурирующую**) гипотезу **H1**.

Основной способ проверки статистической гипотезы – вычислить по имеющейся выборке значение некоторой случайной величины, закон распределения которой известен. Отсюда появляется понятие **статистического критерия** – случайной величины K, имеющей известный закон распределения и позволяющей проверить гипотезу H0. **Областью принятия гипотезы** называют совокупность значений K, при которых нулевая гипотеза принимается, а критической областью – при которых отвергается.

$K = \{w < x1\} \cup \{w < x2\}$,

где w – числовая функция (статистика критерия), x1 и x2 – критические значения статистики, при которых гипотеза принимается.

В результате применения статистического критерия возможно возникновение двух видов ошибки:

- **Ошибка первого рода** – верная гипотеза H0 отвергается.

  Вероятность возникновения такого вида ошибки называется **уровнем значимости** α.

- **Ошибка второго рода** – верная гипотеза H1 отвергается.

Вероятность возникновения такого вида ошибки – β. Величина 1-β определяет **мощность критерия**.

Чем меньше мощность критерия, тем больше β, следовательно при выборе уровня значимости критическую область необходимо стоить так, чтобы мощность критерия была как можно больше.

При осуществлении проверки статистической гипотезы путем построения критической области для значений статистики критерия и его сравнения с наблюдаемым значением статистики нам приходится сталкиваться с некоторым неудобством: критическое значение связано с данным уровнем значимости, а значит, когда он получится другим, вычислять критическое значение придется заново. Чтобы обойти вычисление критических значений для каждого уровня значимости и упростить при этом принятие решения о выполнении гипотезы было введено понятие P-значения. Этот способ проверки заключен в оценивании на основе имеющихся данных вероятности того, что нулевая гипотеза имеет право на существование.

P-value – достигаемый уровень значимости – значение статистического критерия, при котором H0 все еще принимается для любого уровня значимости α.

$$PV = \max\{\alpha \in [0,1] \mid t(x) > c_\alpha\}$$

То есть, если PV(x) <α – гипотеза H0 отвергается, а при PV(x) ≥ α – гипотеза H0 принимается.

2.3. Основной критерий Шапиро – Уилка

Для проверки гипотезы о распределении непрерывной случайной величины применяются статистические критерии двух видов: общие критерии согласия, используемые при проверке гипотезы о согласии реального распределения с каким-либо предполагаемым теоретическим – равномерным, экспоненциальным и т. д. (например омега-квадрат, Андерсона-Дарлинга, Колмогорова), а также специальные, оценивающие согласие реального распределения с определенным видом теоретического.

В последние годы для проверки нормальности того или иного распределения общественность начала отдавать предпочтение специальному критерию Шапиро – Уилка, основанному на оптимальной несмещенной линейной оценке дисперсии к её стандартной оценке методом максимального подобия. Тогда гипотеза H0 звучит так:

«случайная величина X распределена нормально». При использовании критерия результаты испытаний располагают в вариационный ряд и рассчитывают значения:

$$nm_2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad b = \sum_{i=1}^{k} a_{n-i+1}(x_{n-i+1} - x_i)$$

где i – номер элемента в вариационном ряду. При этом, если n чётное, k=n/2, если n нечётное, k=(n-1)/2, а значения $an-i+1$ – табличные величины.

Таблица значений для малых i.

i	a_{n-i+1}
1	$(0{,}0081356n^4 - 1{,}3596n^3 + 87{,}592n^2 - 2808{,}2n + 78028)/100000$
2	$(0{,}0005642n^5 - 0{,}096475n^4 + 6{,}418n^3 - 204{,}59n^2 + 2849{,}1n + 19225)/100000$
3	$(-0{,}000053n^6 + 0{,}010464n^5 - 0{,}83717n^4 + 35{,}172n^3 - 823{,}97n^2 + 10190n - 26059)/100000$
4	$(-0{,}00008785n^6 + 0{,}017143n^5 - 1{,}3644n^4 + 56{,}8921n^3 - 1321{,}67n^2 + 16417{,}8n - 64907)/100000$
5	$(-0{,}0000637n^6 + 0{,}012953n^5 - 1{,}08323n^4 + 47{,}9523n^3 - 1197{,}88n^2 + 16280{,}8n - 77227)/100000$
6	$(0{,}001213n^5 - 0{,}22039n^4 + 15{,}932n^3 - 578{,}01n^2 + 10675{,}3n - 64930)/100000$
7	$(0{,}001058n^5 - 0{,}19846n^4 + 14{,}8811n^3 - 563{,}328n^2 + 10954n - 74246)/100000$
8	$(0{,}0009663n^5 - 0{,}18425n^4 + 14{,}2448n^3 - 558{,}464n^2 + 11321{,}7n - 83480)/100000$
9	$(0{,}000936n^5 - 0{,}18321n^4 + 14{,}431n^3 - 578{,}383n^2 + 12047{,}5n - 94506)/100000$
10	$(-0{,}021445n^4 + 3{,}5688n^3 - 227{,}115n^2 + 6687n - 66534)/100000$
11	$(-0{,}01937n^4 + 3{,}3178n^3 - 218{,}207n^2 + 6675n - 70767)/100000$
12	$(-0{,}01757n^4 + 3{,}0973n^3 - 210{,}36n^2 + 6671{,}5n - 74844)/100000$
13	$(-0{,}01577n^4 + 2{,}8668n^3 - 201{,}302n^2 + 6621{,}8n - 78311)/100000$
14	$(0{,}4448n^3 - 64{,}902n^2 + 3325n - 51098)/100000$
15	$(0{,}4227n^3 - 63{,}247n^2 + 3332{,}2n - 53673)/100000$
16	$(0{,}4046n^3 - 61{,}999n^2 + 3353{,}2n - 56378)/100000$
17	(0,3853n3 - 60,444n^2 + 3354,8n - 58703)/100000
18	$(0{,}3532n^3 - 57{,}207n^2 + 3282{,}5n - 59931)/100000$
19	$(-11{,}224n^2 + 1322{,}1n - 33480)/100000$
20	$(-11{,}072n^2 + 1331{,}1n - 35023)/100000$
21	$(-10{,}898n^2 + 1337{,}8n - 36508)/100000$
22	$(-10{,}833n^2 + 1354{,}4n - 38200)/100000$
23	$(-10{,}714n^2 + 1365{,}6n - 39754)/100000$
24	(335n - 15708)/100000
25	0,0035

Критерий будет надежен при n ∈ [8;50].

Для расчета статистики критерия применяют формулу W =b2/nm2 , после чего найденное значение W сравнивают с табличным и, если вычисленная величина будет меньше табличной, H0 отклоняется и распределение не считают нормальным.

В данной работе критерий Шапиро – Уилка будет реализован с помощью Python благодаря функции shapiro(x), которая принадлежит библиотеке scipy.stats. Эта функция принимает на вход массив выборочных данных и возвращает статистику критерия W и p-значение для проверки гипотезы.

2.4. Вспомогательный критерий Колмогорова

Выбранный критерий рассматривает максимальную абсолютную величину разности значений эмпирической и соответствующей ей теоретической функций распределения.

$$D = \max \left| F_n(x) - F(x) \right|$$

В представленной работе будет находиться разность между эмпирической функцией распределения p-values и равномерным непрерывным распределением на отрезке [0;1].

Были проведены доказательства, подтверждающие, что вне зависимости от функции распределения F(x), вероятность неравенства $P(D\sqrt{(n \geqslant \lambda)})$ (при неограниченном n) стремится к пределу:

$$P(\lambda) = 1 - \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2\lambda^2}$$

Таким образом, задав уровень значимости α, можно найти соответствующее критическое значение λ из соотношения $P(\lambda) = \alpha$.

3. Предварительный анализ данных

3.1. Количество торговых дней

Как было сказано ранее, в работе рассмотрены акции, входящие в индекс «DAX». Список компаний и соответствующих тикеров представлен в «Таблице 1».

Таблица 1. Тикеры компаний.

Тикер	Компания
BMW	Bmw
BAYN	Bayer
DAI	Daimler
ADS	Adidas
VOW	Volkswagen
SIE	Siemens
SAP	SAP
HEN	Henkel
HEI	HeidelbergCement
EOAN	E.ON

Начнем работу с проверки гипотезы для данных, выгруженных с периодичностью 1 день и интервалом в 5 лет. Построим таблицу, отображающую количество торговых дней для каждой компании индекса, чтобы исключить (если имеются) тикеры компаний, торговавших на рынке на протяжение недолгого времени.

Таблица 2. Количество торговых дней.

YEAR	2016	2017	2018	2019	2020
TICKER					
ADS	255	252	239	249	253
BAYN	255	252	239	249	253
BMW	255	252	249	249	253
DAI	255	252	241	249	253
EOAN	255	252	241	249	252
HEI	255	252	241	249	252
HEN	255	252	241	249	252
SAP	255	252	241	249	252
SIE	255	252	241	249	252
VOW	255	252	241	249	234

Исходя из таблицы можно сделать вывод, что все компании торговали акциями в течение времени, достаточного для последующего исследования логарифмических доходностей, а значит нам не требуется исключать из списка какие-либо данные.

3.2. Максимальное относительное изменение цен

Теперь необходимо проверить наличие резких скачков цен на акции (вверх или вниз). Если цена изменится на 50 и более процентов, то тикер соответствующей компании будет удален из перечня исследуемых.

Чтобы определить максимальные и минимальные ценовые скачки необходимо воспользоваться данными поля «CLOSE»: находим отношение цены на сегодняшний день к цене за предыдущий и выражаем в процентах.

Таблица 3. Максимальные дневные ценовые скачки, %

	Тикер	2016	2017	2018	2019	2020
0	BMW	4.74	2.99	4.85	4.07	12.54
1	BAYN	5.16	4.21	5.22	9.05	7.77
2	DAI	4.44	3.21	4.30	4.83	26.05
3	ADS	6.25	9.17	11.62	8.34	8.49
4	VOW	6.88	5.08	5.10	4.81	8.81
5	SIE	8.62	5.63	6.53	5.08	9.97
6	SAP	5.68	2.95	5.02	11.97	8.04
7	HEN	5.04	4.19	4.14	3.58	6.74
8	HEI	5.22	6.33	4.13	4.64	11.35
9	EOAN	8.40	5.29	5.55	3.85	7.11

Таблица 4. Минимальные дневные ценовые скачки, %

	Тикер	2016	2017	2018	2019	2020
0	BMW	-7.53	-3.20	-5.60	-5.23	-12.08
1	BAYN	-8.20	-4.10	-10.70	-9.93	-12.15
2	DAI	-8.22	-4.12	-6.26	-7.22	-16.35
3	ADS	-6.28	-4.66	-7.03	-5.37	-12.97
4	VOW	-10.00	-3.81	-4.66	-4.53	-11.86
5	SIE	-7.41	-3.73	-4.94	-4.12	-11.31
6	SAP	-5.63	-3.46	-6.08	-5.37	-22.37
7	HEN	-4.11	-4.15	-3.60	-10.14	-7.10
8	HEI	-8.03	-3.89	-8.86	-4.38	-12.70
9	EOAN	-14.77	-4.61	-4.06	-5.68	-12.29

Полученные данные наглядно демонстрируют, что максимальные скачки цен на акции произошли в 2020 году: максимальный рост (на 26,05%) зафиксирован в компании Daimler, а максимальное падение – у компании SAP (на 22,37%).

Построим для перечисленных компаний графики цен акций.

Рисунок 1. График изменения цен компании Daimler.

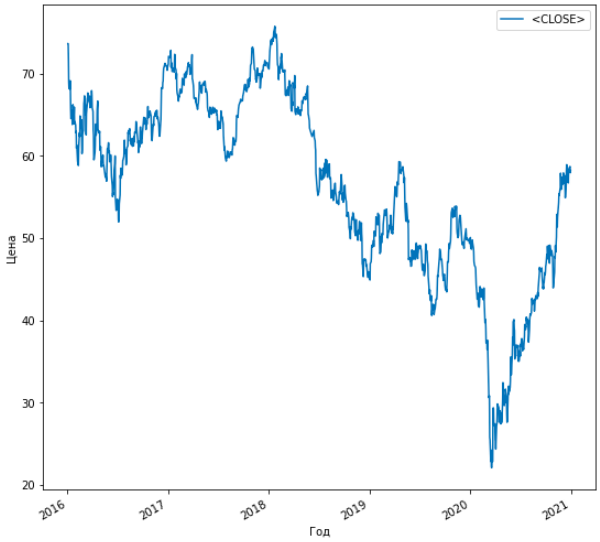
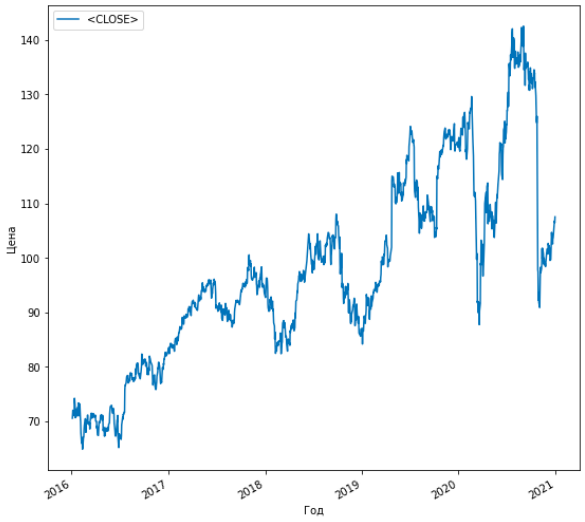


Рисунок 2. График изменения цен компании SAP.



Проведя предварительный анализ данных, мы обнаружили, что все компании пригодны для качественной проверки гипотезы о нормальном распределении логарифмической доходности по выбранному критерию, ни один тикер не был исключен, ведь даже скачки цен оказались незначительными, а значит можно смело приступать к практической части работы.

4. Практическая часть

4.1 Проверка гипотезы на реальных данных (за 5 лет)

На предыдущих этапах курсовой работы удалось убедиться, что выбранный критерий рационален в использовании и алгоритм работает. Аналогично процедуре проверки модельных данных я произвела расчет логарифмической доходности всех компаний за 5 лет.

Таблица 5. Логарифмические доходности каждой компании.

DATE	BMW	BAYN	DAI	ADS	VOW	SIE	SAP	HEN	HEI	EOAN
2016-01-04	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2016-01-05	-0.004672	0.001799	-0.000136	-0.003433	-0.040361	0.006294	0.011971	-0.003036	-0.003915	0.010794
2016-01-06	-0.033669	-0.014486	-0.025725	-0.011645	-0.020808	-0.003958	0.008642	-0.005080	-0.018705	-0.004019
2016-01-07	-0.038347	-0.025867	-0.039221	-0.015662	-0.033351	-0.019315	-0.009903	-0.022350	-0.021734	-0.028597
2016-01-08	-0.023687	-0.027997	-0.011809	-0.017109	0.000869	-0.006921	-0.004073	-0.007423	-0.022793	-0.019074
...	...	...	...	...	...	...	...	...	...	...
2020-12-22	-0.009650	0.013048	-0.005452	0.010626	-0.008287	-0.005408	0.010477	0.007122	0.022503	-0.001135
2020-12-23	0.017303	0.001968	0.027652	0.002044	0.020721	0.020773	0.000386	0.003984	0.010093	0.013989
2020-12-28	-0.004093	0.007424	0.007008	0.015530	0.007857	0.019174	0.029467	0.015562	0.001618	0.021282
2020-12-29	-0.002190	-0.002984	-0.013030	-0.002684	-0.009040	-0.018832	-0.002814	0.001304	0.000323	-0.003516
2020-12-30	-0.004394	-0.001134	0.004820	0.003688	0.000395	0.009545	0.009908	-0.000217	-0.002266	0.001320

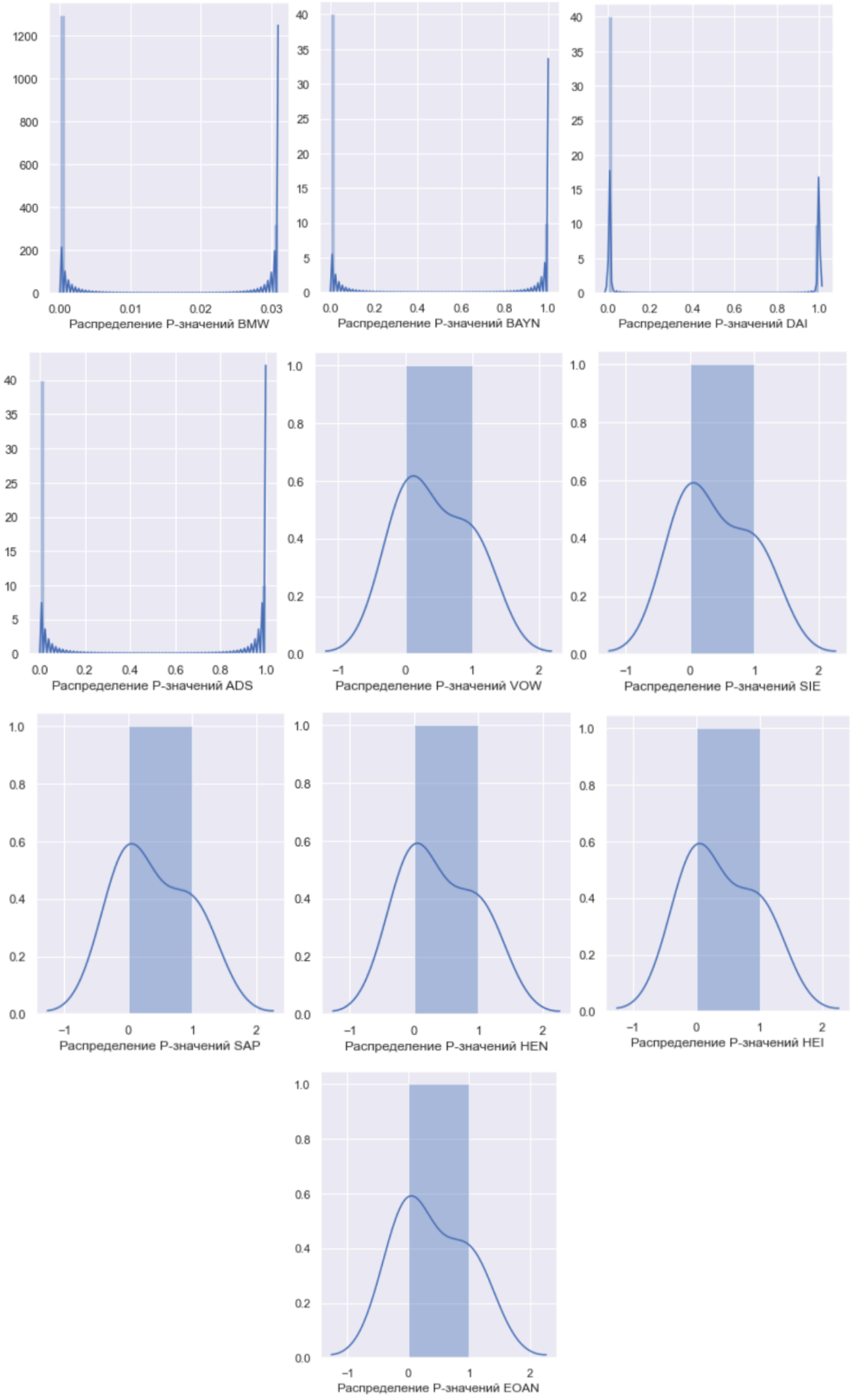
Агрегируем результаты по годам для каждого тикера и вычислим p-значения статистики Шапиро-Уилка.

Таблица 6. P-значения статистики Шапиро-Уилка для каждой компании.

DATE	BMW	BAYN	DAI	ADS	VOW	SIE	SAP	HEN	HEI	EOAN
2016	0.030953	0.000000	0.000014	0.000012	0.017807	0.000000	0.000017	0.000243	0.000004	0.000000
2017	0.000273	0.000704	0.000005	0.000000	0.000047	0.000000	0.000995	0.000000	0.000001	0.000898
2018	0.000277	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
2019	0.000283	0.000000	0.008533	0.000001	0.191476	0.000147	0.000000	0.000000	0.007137	0.000010
2020	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Полученные величины в большинстве своем очень малы. Лишь в исключительных случаях можно наблюдать значение больше уровня значимости в 5% (ячейки в Таблице 9, отмеченные синим цветом). Это значит, что вероятность отклонения нулевой гипотезы H0 приближена к 100%. Подтвердим это предположение построив гистограммы.

Рисунки 3.1–3.10. P-значения статистики Шапиро-Уилка для каждой компании.



Действительно, ни о какой равномерности сказать нельзя. Отсюда следует вывод о том, что объем выборки слишком большой и необходимо рассмотреть более узкий временной период.

4.2. Проверка гипотезы на меньших объемах выборки

В связи с последними событиями (пандемией коронавируса) я решила перейти к сужению временного интервала и рассмотреть следующие периоды: до начала пандемии коронавируса (февраль-май 2019г), пик распространения (февраль-май 2020г) и спад (сентябрь-декабрь 2020г). Проанализируем логдоходности в излеченные периоды и построим таблицы p-values. Отметим на таблице синим цветом значения, которые больше α = 0,05.

Таблица 7. P-значения Шапиро-Уилка февраль-май 2019г.

DATE	BMW	BAYN	DAI	ADS	VOW	SIE	SAP	HEN	HEI	EOAN
2	0.049723	0.667761	0.321989	0.394677	0.483831	0.836473	0.321353	0.676269	0.064457	0.966832
3	0.228770	0.001680	0.998634	0.288508	0.849867	0.002223	0.825527	0.530610	0.283643	0.049015
4	0.374116	0.028421	0.135918	0.742271	0.922778	0.488282	0.000002	0.509341	0.194387	0.573394
5	0.042993	0.849979	0.121977	0.026855	0.086580	0.079097	0.845146	0.007246	0.595313	0.000178

Таблица 8. P-значения Шапиро-Уилка февраль-май 2020г.

DATE	BMW	BAYN	DAI	ADS	VOW	SIE	SAP	HEN	HEI	EOAN
2	0.912962	0.439886	0.901352	0.343450	0.785974	0.030409	0.329773	0.784351	0.265244	0.257976
3	0.046619	0.535190	0.008237	0.450518	1.000000	0.464296	0.364032	0.718024	0.864582	0.316613
4	0.783321	0.699506	0.863711	0.231352	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
5	0.721745	0.043428	0.603926	0.391978	0.663933	0.962786	0.811495	0.439434	0.881789	0.186947

Таблица 9. P-значения Шапиро-Уилка сентябрь-декабрь 2020г.

DATE	BMW	BAYN	DAI	ADS	VOW	SIE	SAP	HEN	HEI	EOAN
9	0.242858	0.105174	0.653373	0.420363	0.179835	0.779003	0.914617	0.979612	0.002085	0.158511
10	0.432032	0.070274	0.989024	0.325534	0.611788	0.004061	0.000000	0.116861	0.905252	0.639659
11	0.456942	0.391506	0.984132	0.649662	0.270473	0.063621	0.577697	0.333525	0.266074	0.841223
12	0.355814	0.038047	0.497853	0.034690	0.008050	0.426334	0.776695	0.934518	0.528852	0.424405

Анализ результатов вычислений позволяет сделать вывод: при существенно меньшем объеме выборки p-значения необходимого уровня встречаются

регулярнее, а значит в большинстве случаев гипотеза о нормальном распределении логарифмической доходности принимается.

Для формулировки качественного вывода вычислим частотные характеристики, показывающие процент принятия гипотезы H0 в каждом периоде по индексу в целом.

Таблица 10. Частотные характеристики принятия гипотезы по критерию Шапиро-Уилка.

```
Доля принятия гипотезы для 1 периода = 75.0 %
Доля принятия гипотезы для 2 периода = 90.0 %
Доля принятия гипотезы для 3 периода = 85.0 %
```

Оказалось, что в период, когда коронавирус еще не имел никакого влияния на экономику, данные распределены менее нормально, чем в последующие периоды. Это может объясняться тем, что в период коронавируса наблюдалась хаотичная ситуация на рынке и постоянные скачки акций вверх и вниз могли в итоге показать большую нормальность, что как раз как раз реже встречается в обычное время, о чем нам говорили работы прошлый лет и мое первое исследование за 5 лет.

Однако подчеркну, что на всех этапах гипотеза H0 все же принимается, поскольку p-значения в большинстве своем больше уровня значимости и процент принятия гипотезы превышает 75%.

4.3. Альтернативная проверка гипотезы на реальных данных критерием Колмогорова-Смирнова

Мне стало интересно сравнить что же покажет аналогичная проверка данных критерием Колмогорова-Смирнова для этих же трех временных периодов. Так же составляем таблицу p-values статистики Колмогорова-Смирнова.

Таблица 11. P-значения Колмогорова-Смирнова февраль-май 2019г.

DATE	BMW	BAYN	DAI	ADS	VOW	SIE	SAP	HEN	HEI	EOAN
2	0.055522	0.075816	0.061574	0.070814	0.064692	0.056192	0.055010	0.048518	0.072201	0.052375
3	0.032698	0.035341	0.040357	0.039191	0.033606	0.028635	0.028919	0.026535	0.029481	0.027787
4	0.052053	0.056251	0.047446	0.048580	0.055035	0.043722	0.069619	0.055800	0.047021	0.048068
5	0.024608	0.028051	0.022349	0.022258	0.022326	0.021768	0.021029	0.018762	0.022078	0.018656

Таблица 12. P-значения Колмогорова-Смирнова февраль-май 2020г.

DATE	BMW	BAYN	DAI	ADS	VOW	SIE	SAP	HEN	HEI	EOAN
2	0.077255	0.062193	0.078965	0.060411	0.069584	0.048330	0.058947	0.053049	0.053469	0.058597
3	0.072781	0.029346	0.186292	0.037646	0.000000	0.048170	0.034862	0.027841	0.060305	0.029687
4	0.067442	0.045385	0.078012	0.051480	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	0.060913	0.044198	0.073803	0.078491	0.065164	0.052468	0.040064	0.038990	0.058700	0.044923

Таблица 13. P-значения Колмогорова-Смирнова сентябрь-декабрь 2020г.

DATE	BMW	BAYN	DAI	ADS	VOW	SIE	SAP	HEN	HEI	EOAN
9	0.021962	0.022162	0.022965	0.021765	0.028227	0.022723	0.030223	0.021884	0.023973	0.018924
10	0.025805	0.027875	0.032021	0.022784	0.022506	0.017836	0.017824	0.017237	0.030153	0.019178
11	0.032777	0.037212	0.038873	0.055784	0.043617	0.042365	0.031875	0.030534	0.031842	0.035075
12	0.058668	0.069860	0.063455	0.052859	0.069384	0.055148	0.055351	0.053661	0.052392	0.055653

Получившиеся значения позволяют сделать вывод, что критерий Колмогорова-Смирнова также применим для исследования. Однако нельзя не заметить, что величины необходимого уровня встречаются реже и не соответствуют позициям из предыдущего исследования.

Попробуем найти объяснение этому факту посчитав процент принятия гипотезы H0 по индексу.

Таблица 14. Частотные характеристики принятия гипотезы по критерию Колмогорова-Смирнова.

```
Доля принятия гипотезы по Колмогорову-Смирнову для 1 периода = 35.0 %
Доля принятия гипотезы по Колмогорову-Смирнову для 2 периода = 65.0 %
Доля принятия гипотезы по Колмогорову-Смирнову для 3 периода = 28.0 %
```

Частотные характеристики значительно уменьшились по сравнению с исследованием по критерию Шапиро-Уилка (упали почти в 3 раза). Как и на прошлом этапе исследования, значения утверждают, что наиболее приближенным к нормальному является распределение логдоходностей периода расцвета коронавируса. Однако критерий Колмогорова-Смирнова показал обратную ситуацию для первого и третьего временных периодов. Быть может, именно поэтому данный критерий не используют как основной для исследования на нормальность, ведь это общий критерий согласия, более универсальный.

5. Заключение

В ходе курсовой работы мною была проведена проверка гипотезы о нормальности распределения логарифмической доходности акций, входящих в ведущий немецкий фондовый индекс «DAX», по критерию Шапиро – Уилка.

Проверка на модельных данных полностью подтвердила гипотезу H0, однако применение критерия на большом объеме реальных данных показало, что, хотя критерий является очень мощным для проверки нормальности, но, к сожалению, имеет ограниченную применимость. Сужение временного периода доказало практичность именно в малых и средних выборках. Были рассмотрены интересные периоды развития коронавируса. Анализ p-значений привел к выводу, что, в большинстве своем, все акции имели нормальное распределение логдоходностей, однако встречающиеся резкие ценовые скачки не позволяют в полной мере доверять этим результатам и все же не дают возможность предсказывать будущую динамику цен. Это подтвердила и проверка данных критерием Колмогорова-Смирнова.

Главная задача исследования решена: с помощью языка Python созданы все необходимые программы, позволяющие совершить проверку различных данных и выборок акций на нормальность распределения логдоходностей.

6. Список использованной литературы

- Браилов А.В. Лекции по математической статистике. –М.: Финакадемия, 2007. – 172 с. (Дата обращения: 21.04.2021).
- Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. –М.: ФИЗМАТЛИТ, 2006. – 816 с. (Дата обращения: 21.04.2021).
- Shapiro S. S., Wilk M. B.An analysis of variance test for normality. — Biometrika, 1965, 52, No3 — p. 591-611. (Дата обращения: 27.04.2021).
- Солодовников А.С., Бабайцев В.А., Браилов А.В. Математика в экономике: учебник в 3-х ч. Ч. 3. Теория вероятностей и математическая статистика. – М.: Финансы и статистика, 2008. – 464 с.: ил. (Дата обращения: 23.04.2021).

Интернет-ресурсы:

- https://www.finam.ru/profile/

7. Приложение

7.1. Проверка гипотезы для модельных данных

Используем критерий Шапиро-Уилка для проверки гипотезы на модельных данных. Применим программу, которая создает три различные выборки из стандартного нормального распределения объемами, отражающими изменения за год (n1 = 252), полгода (n2 = 126) и квартал (n3 = 63) и проводим серию из 10000 испытаний методом Монте-Карло для каждой из этой выборки. Далее считаем 999 квантилей распределения имеющейся статистики при условии, что выдвинутая гипотеза о нормальном распределении логарифмической доходности H0 верна.

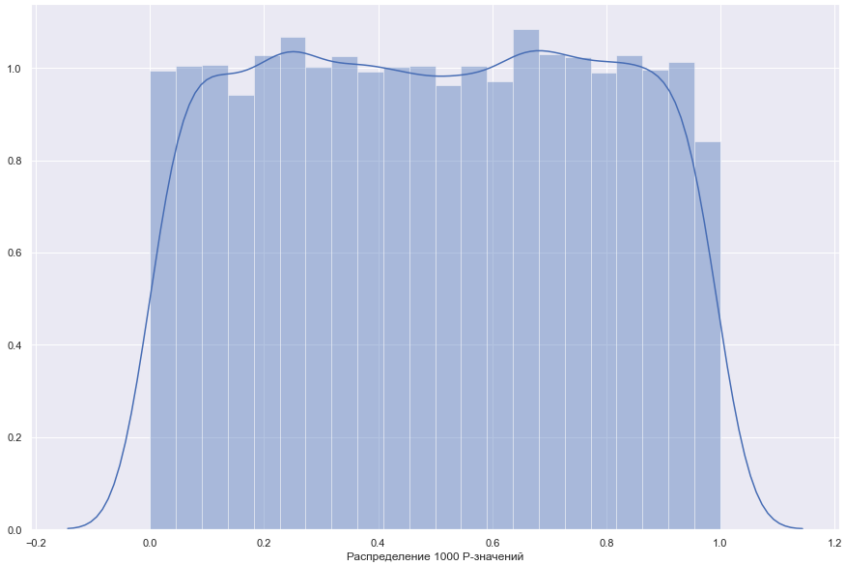
Таблица 15. Квантили уровней 0.1, 0.2, …, 0.9 для трёх выборок.

Level	252	126	63
0.1	0.990593	0.982177	0.968440
0.2	0.992140	0.985340	0.974042
0.3	0.993146	0.987356	0.977762
0.4	0.993890	0.988813	0.980231
0.5	0.994514	0.989958	0.982563
0.6	0.995081	0.991000	0.984478
0.7	0.995597	0.992031	0.986286
0.8	0.996120	0.993014	0.988088
0.9	0.996756	0.994211	0.990280

Стандартная ошибка при n = 252: 0.0026843850156179506
Стандартная ошибка при n = 126: 0.005204045841789566
Стандартная ошибка при n = 63: 0.00957079373629624

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

Рисунок 4. Гистограмма 1000 p-значений модельных данных (на нормально распределенной выборке).



Гистограмма наглядно демонстрирует, что p-значения на заданном отрезке распределены равномерно. Это подтверждает и вычисленное p-значение критерия Колмогорова-Смирнова, равное 0.52731, что больше уровня значимости 0.05.
Следовательно, мы принимаем гипотезу о нормальном распределении и можем перейти к проверке критерия на реальных данных.

7.2. Выбор альтернативной гипотезы и оценка мощности критерия

Для формулировки альтернативных гипотез, с помощью которых можно вычислить мощность критерия Шапиро-Уилка, выберем такие распределения, которые являются наиболее близкими к нормальному: распределение Лапласа, распределение χ2 и распределение Стьюдента (оба с двумя степенями свободы). С помощью соответствующих этим распределениям функциям в Python вычисляем 1000 p-значений для каждого из них, используя различные размерности выборок (n1 = 252, n2 = 126, n3 = 63). Затем рассчитываем мощность критерия при уровне значимости равном 0,05. Результаты записываем в таблицу.

Таблица 16. Мощность критерия.

	name	252	126	63
0	Лаплас	0.992	0.864	0.615
1	Хи квадрат	1.000	1.000	1.000
2	Стьюдент	1.000	0.994	0.929

Исходя из данных, полученных в таблице, следует вывод о том, что выбранный критерий обладает наибольшей эффективностью для распределений Стьюдента и Хи квадрат в двумя степенями свободы. Распределение Лапласа будет эффективно лишь при больших объемах выборки.
Проведем дополнительный анализ, где рассмотрим распределения Стьюдента и Хи квадрат с большими степенями свободы.

Таблица 17.1. Мощность критерия (степень свободы = 3).

	name	252	126	63
0	Хи квадрат	1.000	1.000	0.995
1	Стьюдент	0.996	0.924	0.716

Таблица 17.2. Мощность критерия (степень свободы = 4).

	name	252	126	63
0	Хи квадрат	1.000	1.000	0.991
1	Стьюдент	0.968	0.793	0.519

Следует вывод, что при увеличении числа степеней свободы распределений мощность критерия снижается, что приводит к росту вероятности совершить ошибку второго рода.

7.3. Характеристики компьютера

```
Тип процессора – 8 GHz 2-ядерный процессор Intel Core i5
Тактовая частота – 2200 МГц
Частота системноиˇ шины – 1066 МГц
Объем кэш-памяти второго уровня (L2) – 256 КБ
```

7.4. Коды программ

```
Импортирование библиотек.
Import pandas as pd
import numpy as np
import plotly.graph_objects as go
import plotly.figure_factory as ff
from scipy.stats import norm
from matplotlib import pyplot as plt
import scipy.stats as st
import seaborn as sns
Импортирование акций.
tickers = ['BMW', 'BAYN', 'DAI', 'ADS', 'VOW', 'SIE', 'SAP', 'HEN',
'HEI', 'EOAN']
def get_data(tickers):
    share_data = pd.read_csv(f'{tickers}year.csv')
    share_data['_DATE'] = pd.to_datetime(share_data['<DATE>'],
format='%Y%m%d')
    share_data['<DATE>'] = pd.to_datetime(share_data['<DATE>'],
format='%Y%m%d').dt.year.astype('int')
    return share_data
BMW = get_data(tickers[0])
BAYN = get_data(tickers[1])
DAI = get_data(tickers[2])
ADS = get_data(tickers[3])
VOW = get_data(tickers[4])
SIE = get_data(tickers[5])
SAP = get_data(tickers[6])
HEN = get_data(tickers[7])
HEI = get_data(tickers[8])
EOAN = get_data(tickers[9])
datas=[BMW,BAYN,DAI,ADS,VOW,SIE,SAP,HEN,HEI,EOAN]
for data in datas:
data.columns=['TICKER','PER','YEAR','TIME','OPEN','HIGH','LOW','CLOSE'
,'VOL','DATE']
    data['YEAR'] = data['DATE'].dt.year
    data['LOGRET'] =
np.log(data['CLOSE'].divide(data['CLOSE'].shift(1)))
    data.fillna(0,inplace=True)
shares =
pd.concat((BMW,BAYN,DAI,ADS,VOW,SIE,SAP,HEN,HEI,EOAN),ignore_index=Tru
e,sort=False)
shares.fillna(0,inplace=True)
Таблица 2. Количество торговых дней.
shares.pivot_table(index='TICKER', columns='YEAR', values='CLOSE',
aggfunc='count')
Таблица 3. Максимальные дневные ценовые скачки, %
def rost(year,ticker):
    df = ticker
```

```
    years = []
    for i in range(len(df['YEAR'])):
        years.append(int((str(df['YEAR'][i])) [:4]))
    df['YEAR'] = years
    ind = list(np.where(df['YEAR'] == year)[0])[:-1]
    pros = [round(df['CLOSE'][i+1]/df['CLOSE'][i]*100-100, 2 ) for i
in ind]
    return max(pros)
akcii = [BMW, BAYN, DAI, ADS, VOW, SIE, SAP, HEN, HEI, EOAN]
tickers = ['BMW', 'BAYN', 'DAI', 'ADS', 'VOW', 'SIE', 'SAP', 'HEN',
'HEI', 'EOAN']
years = range(2016,2021)
max_rost = pd.DataFrame()
max_rost['Тикер'] = tickers
for year in years:
    changes_max = []
    for a in akcii:
        changes_max.append(rost(year, a))
    max_rost[str(year)] = changes_max
max_rost
```

Таблица 4. Минимальные дневные ценовые скачки, %

```
def padenie(year,ticker):
    df = ticker
    years = []
    for i in range(len(df['YEAR'])):
        years.append(int((str(df['YEAR'][i]))[:4]))
    df['YEAR'] = years
    ind = list(np.where(df['YEAR'] == year)[0])[:-1]
    pros = [round(df['CLOSE'][i+1]/df['CLOSE'][i]*100-100, 2) for i in
ind]
    return min(pros)
min_padenie = pd.DataFrame()
min_padenie['Тикер'] = tickers
for year in years:
    changes = []
    for a in akcii:
        changes.append(padenie(year, a))
    min_padenie[str(year)] = changes
min_padenie
```

Рисунок 1. График изменения цен компании Daimler.

```
shares[shares['TICKER'] == 'DAI'].plot(x='DATE', y='CLOSE',
figsize=(9,9))
plt.xlabel('Год')
plt.ylabel('Цена')
```

Рисунок 2. График изменения цен компании SAP.

```
shares[shares['TICKER'] == 'SAP'].plot(x='DATE', y='CLOSE',
figsize=(9,9))
plt.xlabel('Год')
```

```
plt.ylabel('Цена')
```
Таблица 5. Логарифмические доходности каждой компании.
```
BMW_l =
pd.DataFrame(index=BMW["DATE"],data={"DATE":BMW["DATE"].values,"BMW":B
MW['LOGRET'].values})
BMW_l.drop_duplicates(inplace=True)
BAYN_l =
pd.DataFrame(index=BAYN["DATE"],data={"DATE":BAYN["DATE"].values,"BAYN
":BAYN['LOGRET'].values})
BAYN_l.drop_duplicates(inplace=True)
DAI_l =
pd.DataFrame(index=DAI["DATE"],data={"DATE":DAI["DATE"].values,"DAI":D
AI['LOGRET'].values})
DAI_l.drop_duplicates(inplace=True)
ADS_l =
pd.DataFrame(index=ADS["DATE"],data={"DATE":ADS["DATE"].values,"ADS":A
DS['LOGRET'].values})
ADS_l.drop_duplicates(inplace=True)
VOW_l =
pd.DataFrame(index=VOW["DATE"],data={"DATE":VOW["DATE"].values,"VOW":V
OW['LOGRET'].values})
VOW_l.drop_duplicates(inplace=True)
SIE_l =
pd.DataFrame(index=SIE["DATE"],data={"DATE":SIE["DATE"].values,"SIE":S
IE['LOGRET'].values})
SIE_l.drop_duplicates(inplace=True)
SAP_l =
pd.DataFrame(index=SAP["DATE"],data={"DATE":SAP["DATE"].values,"SAP":S
AP['LOGRET'].values})
SAP_l.drop_duplicates(inplace=True)
HEN_l =
pd.DataFrame(index=HEN["DATE"],data={"DATE":HEN["DATE"].values,"HEN":H
EN['LOGRET'].values})
HEN_l.drop_duplicates(inplace=True)
HEI_l =
pd.DataFrame(index=HEI["DATE"],data={"DATE":HEI["DATE"].values,"HEI":H
EI['LOGRET'].values})
HEI_l.drop_duplicates(inplace=True)
EOAN_l =
pd.DataFrame(index=EOAN["DATE"],data={"DATE":EOAN["DATE"].values,"EOAN
":EOAN['LOGRET'].values})
EOAN_l.drop_duplicates(inplace=True)
BMW_l.fillna(0,inplace=True)
BAYN_l.fillna(0,inplace=True)
DAI_l.fillna(0,inplace=True)
ADS_l.fillna(0,inplace=True)
VOW_l.fillna(0,inplace=True)
SIE_l.fillna(0,inplace=True)
```

```
SAP_l.fillna(0,inplace=True)
HEN_l.fillna(0,inplace=True)
HEI_l.fillna(0,inplace=True)
EOAN_l.fillna(0,inplace=True)
shares_logs=pd.DataFrame(index=BMW_l['DATE'],
                         data={"BMW":BMW_l["BMW"],
                         "BAYN":BAYN_l["BAYN"],
                         "DAI":DAI_l["DAI"],
                         "ADS":ADS_l["ADS"],
                         "VOW":VOW_l["VOW"],
                         "SIE":SIE_l["SIE"],
                         "SAP":SAP_l["SAP"],
                         "HEN":HEN_l["HEN"],
                         "HEI":HEI_l["HEI"],
                         "EOAN":EOAN_l["EOAN"]})
shares_logs
```

Таблица 6. P-значения статистики Шапиро-Уилка для каждой компании.

```
shares_year = shares_logs.groupby(shares_logs.index.year).agg({
                           'BMW': lambda x: st.shapiro(x)[1],
                           'BAYN': lambda x: st.shapiro(x)[1],
                           'DAI': lambda x: st.shapiro(x)[1],
                           'ADS': lambda x: st.shapiro(x)[1],
                           'VOW': lambda x: st.shapiro(x)[1],
                           'SIE': lambda x: st.shapiro(x)[1],
                           'SAP': lambda x: st.shapiro(x)[1],
                           'HEN': lambda x: st.shapiro(x)[1],
                           'HEI': lambda x: st.shapiro(x)[1],
                           'EOAN': lambda x: st.shapiro(x)[1]})
shares_year
```

Рисунки 3.1-3.10. P-значения статистики Шапиро-Уилка для каждой компании.

```
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.BMW, axlabel = 'Распределение P-значений
BMW')
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.BAYN, axlabel = 'Распределение P-значений
BAYN')
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.DAI, axlabel = 'Распределение P-значений
DAI')
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.ADS, axlabel = 'Распределение P-значений
ADS')
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.VOW, axlabel = 'Распределение P-значений
VOW')
sns.set(rc={'figure.figsize':(4, 5)})
```

```
sns.distplot(shares_year.SIE, axlabel = 'Распределение P-значений
SIE')
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.SAP, axlabel = 'Распределение P-значений
SAP')
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.HEN, axlabel = 'Распределение P-значений
HEN')
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.HEI, axlabel = 'Распределение P-значений
HEI')
sns.set(rc={'figure.figsize':(4, 5)})
sns.distplot(shares_year.EOAN, axlabel = 'Распределение P-значений
EOAN')
```

Таблица 7. P-значения Шапиро-Уилка февраль-май 2019г.

```
shares_2019 = shares_logs[(shares_logs.index.year == 2019)]
shares_2019m_ = shares_2019[(shares_2019.index.month >= 2)]
shares_2019m = shares_2019m_[(shares_2019m_.index.month <= 5)]
shares_months_2019 =
shares_2019m.groupby(shares_2019m.index.month).agg({
                            'BMW': lambda x: st.shapiro(x)[1],
                            'BAYN': lambda x: st.shapiro(x)[1],
                            'DAI': lambda x: st.shapiro(x)[1],
                            'ADS': lambda x: st.shapiro(x)[1],
                            'VOW': lambda x: st.shapiro(x)[1],
                            'SIE': lambda x: st.shapiro(x)[1],
                            'SAP': lambda x: st.shapiro(x)[1],
                            'HEN': lambda x: st.shapiro(x)[1],
                            'HEI': lambda x: st.shapiro(x)[1],
                            'EOAN': lambda x: st.shapiro(x)[1]})
shares_months_2019.style.apply(check_005)
```

Таблица 8. P-значения Шапиро-Уилка февраль-май 2020г.

```
shares_2020 = shares_logs[(shares_logs.index.year == 2020)]
shares_2020m_ = shares_2020[(shares_2020.index.month >= 2)]
shares_2020m = shares_2020m_[(shares_2020m_.index.month <= 5)]
shares_months_2020 =
shares_2020m.groupby(shares_2020m.index.month).agg({
                            'BMW': lambda x: st.shapiro(x)[1],
                            'BAYN': lambda x: st.shapiro(x)[1],
                            'DAI': lambda x: st.shapiro(x)[1],
                            'ADS': lambda x: st.shapiro(x)[1],
                            'VOW': lambda x: st.shapiro(x)[1],
                            'SIE': lambda x: st.shapiro(x)[1],
                            'SAP': lambda x: st.shapiro(x)[1],
                            'HEN': lambda x: st.shapiro(x)[1],
                            'HEI': lambda x: st.shapiro(x)[1],
                            'EOAN': lambda x: st.shapiro(x)[1]})
shares_months_2020.style.apply(check_005)
```

Таблица 9. P-значения Шапиро-Уилка сентябрь-декабрь 2020г.

```
shares_2020m__ = shares_2020[(shares_2020.index.month >= 9)]
shares_months_2020__ =
shares_2020m__.groupby(shares_2020m__.index.month).agg({
                            'BMW': lambda x: st.shapiro(x)[1],
                            'BAYN': lambda x: st.shapiro(x)[1],
                            'DAI': lambda x: st.shapiro(x)[1],
                            'ADS': lambda x: st.shapiro(x)[1],
                            'VOW': lambda x: st.shapiro(x)[1],
                            'SIE': lambda x: st.shapiro(x)[1],
                            'SAP': lambda x: st.shapiro(x)[1],
                            'HEN': lambda x: st.shapiro(x)[1],
                            'HEI': lambda x: st.shapiro(x)[1],
                            'EOAN': lambda x: st.shapiro(x)[1]})
shares_months_2020__.style.apply(check_005)
```

Таблица 10. Частотные характеристики принятия гипотезы по критерию Шапиро-Уилка.

```
def per1 (data,a,b):
    tickers = ['BMW', 'BAYN', 'DAI', 'ADS', 'VOW', 'SIE', 'SAP',
'HEN', 'HEI', 'EOAN']
    k_per = 0
    for i in range (a,b+1):
        for k in tickers:
            if data[k][i] >= 0.05:
                k_per += 1

    chisl = int(k_per)
    znam = len(tickers)*(b-a+1)
    return chisl/znam*100
print('Доля принятия гипотезы для 1 периода
=',per1(shares_months_2019,2,5),'%')
print('Доля принятия гипотезы для 2 периода
=',per1(shares_months_2020,2,5),'%')
print('Доля принятия гипотезы для 3 периода
=',per1(shares_months_2020__,9,12),'%')
```

Таблица 11. P-значения Колмогорова-Смирнова февраль-май 2019г.

```
shares_months_2019ksss =
shares_2019m.groupby(shares_2019m.index.month).agg({
                            'BMW': lambda x: st.kstest(x, 'norm')[1],
                            'BAYN': lambda x: st.kstest(x, 'norm')[1],
                            'DAI': lambda x: st.kstest(x, 'norm')[1],
                            'ADS': lambda x: st.kstest(x, 'norm')[1],
                            'VOW': lambda x: st.kstest(x, 'norm')[1],
                            'SIE': lambda x: st.kstest(x, 'norm')[1],
                            'SAP': lambda x: st.kstest(x, 'norm')[1],
                            'HEN': lambda x: st.kstest(x, 'norm')[1],
                            'HEI': lambda x: st.kstest(x, 'norm')[1],
```

```
                            'EOAN': lambda x: st.kstest(x,
'norm')[1]})
shares_months_2019ksss.style.apply(check_005)
```
Таблица 12. P-значения Колмогорова-Смирнова февраль-май 2020г.
```
shares_months_2020kss =
shares_2020m.groupby(shares_2020m.index.month).agg({
                            'BMW': lambda x: st.kstest(x, 'norm')[1],
                            'BAYN': lambda x: st.kstest(x, 'norm')[1],
                            'DAI': lambda x: st.kstest(x, 'norm')[1],
                            'ADS': lambda x: st.kstest(x, 'norm')[1],
                            'VOW': lambda x: st.kstest(x, 'norm')[1],
                            'SIE': lambda x: st.kstest(x, 'norm')[1],
                            'SAP': lambda x: st.kstest(x, 'norm')[1],
                            'HEN': lambda x: st.kstest(x, 'norm')[1],
                            'HEI': lambda x: st.kstest(x, 'norm')[1],
                            'EOAN': lambda x: st.kstest(x,
'norm')[1]})
shares_months_2020kss.style.apply(check_005)
```
Таблица 13. P-значения Колмогорова-Смирнова сентябрь-декабрь 2020г.
```
shares_months_2020__kss =
shares_2020m__.groupby(shares_2020m__.index.month).agg({
                            'BMW': lambda x: st.kstest(x, 'norm')[1],
                            'BAYN': lambda x: st.kstest(x, 'norm')[1],
                            'DAI': lambda x: st.kstest(x, 'norm')[1],
                            'ADS': lambda x: st.kstest(x, 'norm')[1],
                            'VOW': lambda x: st.kstest(x, 'norm')[1],
                            'SIE': lambda x: st.kstest(x, 'norm')[1],
                            'SAP': lambda x: st.kstest(x, 'norm')[1],
                            'HEN': lambda x: st.kstest(x, 'norm')[1],
                            'HEI': lambda x: st.kstest(x, 'norm')[1],
                            'EOAN': lambda x: st.kstest(x,
'norm')[1]})
shares_months_2020__kss.style.apply(check_005)
```
Таблица 14. Частотные характеристики принятия гипотезы по критерию Колмогорова-Смирнова.
```
print('Доля принятия гипотезы по Колмогорову-Смирнову для 1 периода
=',per1(shares_months_2019ksss,2,5),'%')
print('Доля принятия гипотезы по Колмогорову-Смирнову для 2 периода
=',per1(shares_months_2020kss,2,5),'%')
print('Доля принятия гипотезы по Колмогорову-Смирнову для 3 периода
=',round(per1(shares_months_2020__kss,9,12),0),'%')
```
Таблица 15. Квантили уровней 0.1, 0.2, …, 0.9 для трёх выборок.
```
n1 = 252
n2 = 126
n3 = 63
k = 10000
shapir_data252 = []
shapir_data126 = []
```

```
shapir_data63 = []
shapir_data = []
for i in range(k):
    viborka = np.random.normal(0, 1, n1)
    shapir = st.shapiro(viborka)
    shapir_data252.append(shapir[0])
    shapir_data.append(shapir[0])
for i in range(k):
    viborka = np.random.normal(0, 1, n2)
    shapir1 = st.shapiro(viborka)
    shapir_data126.append(shapir1[0])
for i in range(k):
    viborka = np.random.normal(0, 1, n3)
    shapir2 = st.shapiro(viborka)
    shapir_data63.append(shapir2[0])
quantile_252 = pd.DataFrame(index = np.arange(0.1, 1, 0.1), columns =
['252'], data = np.quantile(np.array(shapir_data252),np.arange(0.1, 1,
0.1)))
print('Стандартная ошибка при n = 252:',np.std(shapir_data252))
quantile_126 = pd.DataFrame(index = np.arange(0.1, 1, 0.1), columns =
['126'], data = np.quantile(np.array(shapir_data126),np.arange(0.1, 1,
0.1)))
print('Стандартная ошибка при n = 126:',np.std(shapir_data126))
quantile_63 = pd.DataFrame(index = np.arange(0.1, 1, 0.1), columns =
['63'], data = np.quantile(np.array(shapir_data63),np.arange(0.1, 1,
0.1)))
print('Стандартная ошибка при n = 63:',np.std(shapir_data63))
quantiles= pd.DataFrame(index = np.arange(0.0005, 1, 0.001), columns =
['value'], data = np.quantile(np.array(shapir_data),np.arange(0.0005,
1, 0.001)))
quantiles.index.name = 'quantile'
quantile=pd.concat((quantile_252,quantile_126,quantile_63),sort=False,
axis = 1)
quantile.index.name = 'Level'
quantile
```

Рисунок 4. Гистограмма 1000 p-значений модельных данных (на нормально распределенной выборке).

```
n = 252
m = 10000
dataa = []
for i in range(m):
    viborka = np.random.normal(0, 1, n)
    dataa.append(st.shapiro(viborka)[1])
sns.set(rc={'figure.figsize':(15, 10)})
sns.distplot(dataa, axlabel = 'Распределение 1000 P-значений')
#проверка критерием Колмогорова
kholm = st.kstest(dataa,'uniform')
round(kholm[1],5)
```

Таблица 16. Мощность критерия.

```
n_ = [252, 126, 63]
lap = ['Лаплас']
for n in n_:
    m = 1000
    laplace = []
    ## Формирование выборок из распределения Лапласа размера n
    for i in range(m):
        vib = st.laplace.rvs(loc = 0, scale = 1,size = n)
        laplace.append(st.shapiro(vib)[1])
    ## Вычисление мощности критерия
    power_laplace = 0
    for i in laplace:
        if i < 0.05:
            power_laplace += 1
    lap.append(power_laplace/m)
chi = ['Хи квадрат']
for n in n_:
    m = 1000
    chi2 = []
    ## Формирование выборок из распределения Хи квадрат размера n
    for i in range(m):
        vib = st.chi2.rvs(df = 2, loc = 0, scale = 1,size = n)
        chi2.append(st.shapiro(vib)[1])
    ## Вычисление мощности критерия
    power_chi = 0
    for i in chi2:
        if i < 0.05:
            power_chi += 1
    chi.append(power_chi/m)
stud = ['Стьюдент']
for n in n_:
    m = 1000
    student = []
    ## Формирование выборок из распределения Стьюдента размера n
    for i in range(m):
        vib = st.t.rvs(df = 2, loc = 0, scale = 1,size = n)
        student.append(st.shapiro(vib)[1])
    ## Вычисление мощности критерия
    power_stud = 0
    for i in student:
        if i < 0.05:
            power_stud += 1
    stud.append(power_stud/m)
df = pd.DataFrame([lap, chi, stud], columns = ['name','252', '126',
'63'])
df
```

Таблица 17. Мощность критерия (степень свободы = r).

```
n_ = [252, 126, 63]
chi = ['Хи квадрат']
for n in n_:
    m = 1000
    chi2 = []
    for i in range(m):
        vib = st.chi2.rvs(df = r, loc = 0, scale = 1,size = n)
        chi2.append(st.shapiro(vib)[1])
    power_chi = 0
    for i in chi2:
        if i < 0.05:
            power_chi += 1
    chi.append(power_chi/m)
stud = ['Стьюдент']
for n in n_:
    m = 1000
    student = []
    ## Формирование выборок из распределения Стьюдента размера n
    for i in range(m):
        vib = st.t.rvs(df = r, loc = 0, scale = 1,size = n)
        student.append(st.shapiro(vib)[1])
    ## Вычисление мощности критерия
    power_stud = 0
    for i in student:
        if i < 0.05:
            power_stud += 1
    stud.append(power_stud/m)
df_ = pd.DataFrame([chi, stud], columns = ['name','252', '126', '63'])
```

7.5. Список файлов

Kachulyak_shares.ipynb – файл JupiterNotebook, работа с акциями.
Kachulyak_model_and_power.ipynb – файл JupiterNotebook, модельные данные, альтернативные гипотезы, мощность.
Курсовая_Качуляк_МГ_ПМ19-4.pdf – Печатная работа.
Файлы тикеров («имя тикера»+year.csv) – 10 файлов формата .csv.

7.6. Время работы программ

Программа	**Время работы, с.**
Импортирование акций.	0.1549136638641
Таблица 2. Количество торговых дней.	0.01895499229431
Таблица 3. Максимальные дневные ценовые скачки, %	2.076689958572
Таблица 4. Минимальные дневные ценовые скачки, %	2.000721216201
Рисунок 1. График изменения цен компании Daimler.	0.0542261600494
Рисунок 2. График изменения цен компании SAP.	0.0435609817504
Таблица 5. Логарифмические доходности каждой компании.	0.0671620368957
Таблица 6. P-значения статистики Шапиро-Уилка для каждой компании.	0.5681908130645

Рисунки 3.1–3.10. P-значения статистики Шапиро-Уилка для каждой компании.	2.94675111770629
Таблица 7. P-значения Шапиро-Уилка февраль-май 2019г.	0.06301093101501
Таблица 8. P-значения Шапиро-Уилка февраль-май 2020г.	0.04990005493164
Таблица 9. P-значения Шапиро-Уилка сентябрь-декабрь 2020г.	0.02842378616333
Таблица 10. Частотные характеристики принятия гипотезы по критерию Шапиро-Уилка.	0.006604909896
Таблица 11. P-значения Колмогорова-Смирнова февраль-май 2019г.	0.0862967967987
Таблица 12. P-значения Колмогорова-Смирнова февраль-май 2020г.	0.07900190353393
Таблица 13. P-значения Колмогорова-Смирнова сентябрь-декабрь 2020г.	0.07852602005004
Таблица 14. Частотные характеристики принятия гипотезы по критерию Шапиро-Уилка.	0.014499187469482
Таблица 15. Квантили уровней 0.1, 0.2, …, 0.9 для трёх выборок.	0.94640588760375
Рисунок 4. Гистограмма 1000 p-значений модельных данных (на нормально распределенной выборке).	0.4659922122955
Таблица 16. Мощность критерия.	0.887383222579