Федеральное государственное образовательное бюджетное учреждение высшего образования

«ФИНАНСОВЫЙ УНИВЕРСИТЕТ

ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

(Финансовый университет)

Департамент анализа данных и машинного обучения

нсовых технологий

Департамент анализа данных, принятия решений и фина

**Расчётно - аналитическая работа по предмету "Анализ данных"**

«Проверка гипотезы о нормальном распределении логарифмической доходности по критерию Дэвида-Хартли-Пирсона»

Курсовая работа на тему

Вид исследуемых данных:

Котировки акций компаний, входящих в индекс ММВБ нефти и газа (MICEX O&G)

*Выполнил:*

студент группы ПИ20-5

Громова М.С.

*Руководитель:*

к.ф.-м.н., доцент Пяткина Д.А.

д-р ф.-м. н. Рябов П.Е

.

Москва

2022

Оглавление

[1. Загрузка исходных данных 3](#_Toc104116378)

[2. Вычисление дополнительных признаков для каждой компании 4](#_Toc104116379)

[3. Исследование изменения цен и объемов торгов акций 4](#_Toc104116380)

[4. Исследование логарифмических доходностей (логдоходностей) акций (удаление выбросов) 14](#_Toc104116381)

[5. Интервальные оценки параметров логдоходностей (в предположении нормального распределения), доверительные интервалы 16](#_Toc104116382)

[6. Проверка гипотезы о нормальности логдоходности по критерию Пирсона (хи-квадрат) 17](#_Toc104116383)

[7. Проверка гипотезы о нормальности логдоходности по критерию Колмогорова-Смирнова 19](#_Toc104116384)

[8. Проверка гипотезы о равенстве нулю математического ожидания недельной логдоходности при альтернативное гипотезе о том, что оно больше нуля 23](#_Toc104116385)

[9. Проверка гипотезы о равенстве дисперсий за последний и предпоследний годы (двусторонний F-тест) 23](#_Toc104116386)

[10. Проверка гипотезы о равенстве средних значений логарифмической доходности за последний и предпоследний годы (двусторонний t-тест с информацией о равенстве дисперсий) 24](#_Toc104116387)

[11. Сравнительный анализ логдоходностей трех компаний 24](#_Toc104116388)

[12. Исследование тесноты связи между логдоходностями 29](#_Toc104116389)

# Загрузка исходных данных

Вариант 102



Тикеры для моего варианта:

-ВТБ ао

-ДагСБ ао

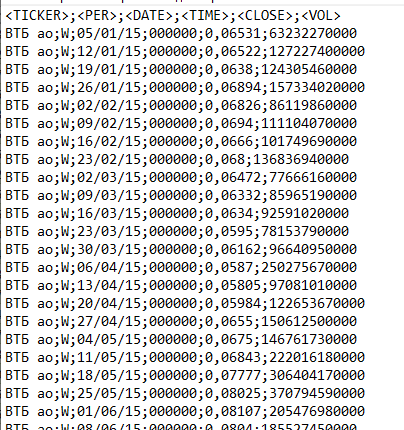
-РОСИНТЕРао

-Соллерс

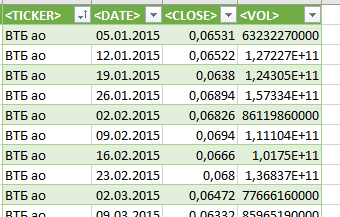
-ФСК ЕЭС ао

Для работы взял первые 3

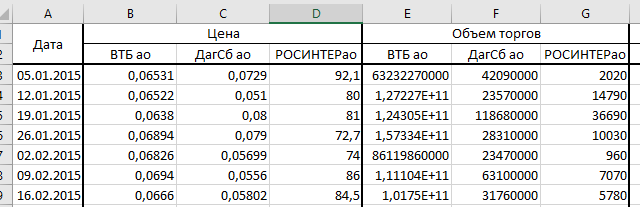
Выгрузили данные по всем 5 Тикерам с сайта mfd.ru за период с 2015 по 2019 год.



Далее загрузили данные в Excel

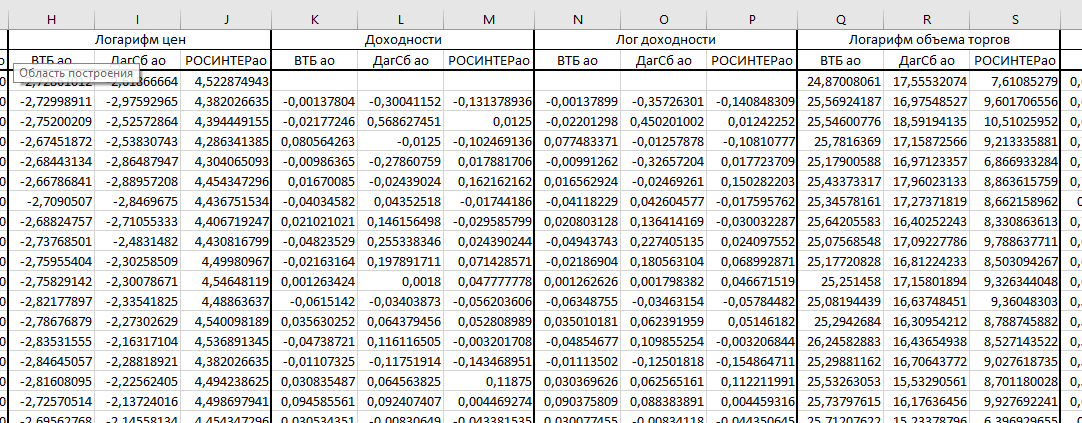


Отредактировали данные по тикерам и расположили следующим образом



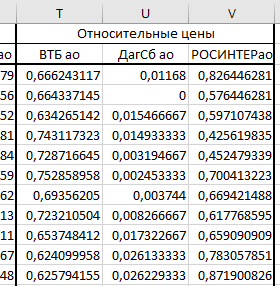
# Вычисление дополнительных признаков для каждой компании

Далее дополнили столбцами Логарифм цены, доходность, логарифм доходности, логарифм объема



# Исследование изменения цен и объемов торгов акций

Для получения представления о поведении различных цен акций создадим дополнительный столбец, в котором вычислим относительные цены по формуле



Теперь мы можем построить график динамики цен всех трех акций сразу.

Здесь можем наблюдать, что цены акций ВТБ ао постепенно снижались, как и РОСИНТЕР ао, а цены ДагСБ ао в 2018 году аномально подскочили.

Для исследования изменения значений показателей удобно будет использовать описательную статистику.

На следующем рисунке представлена описательная статистика по цене акций. Здесь видим, что акции ВТБ самые дешевые, а акции РОСИНТЕР в разы дороже остальных.

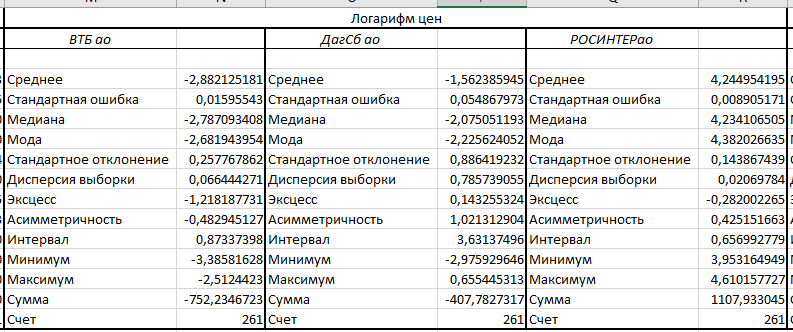


Далее описательная статистика по объему торгов



Наибольший объем торгов у самых дешевых акций (ВТБ).

Далее описательная статистика логарифма цен. Видим, что наибольший разброс значений имеет ДагСБ ао.

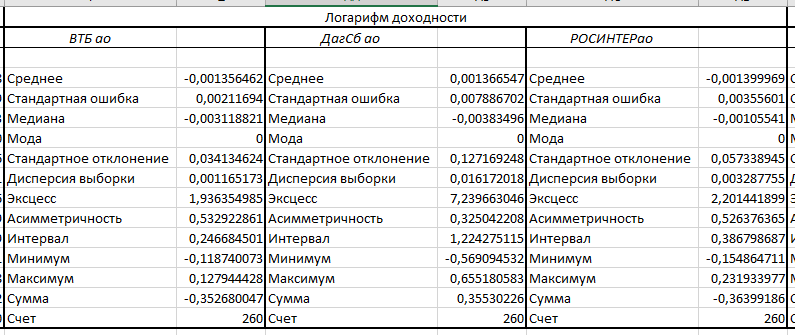


Далее описательная статистика доходности.

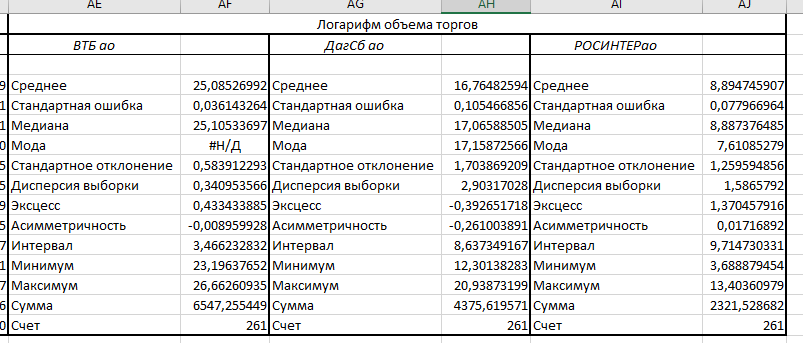


Здесь можно наблюдать то, что видели изначально на динамике цен, доходность акций ВТБ ао на самом деле является убытком, то есть отрицательна в среднем.

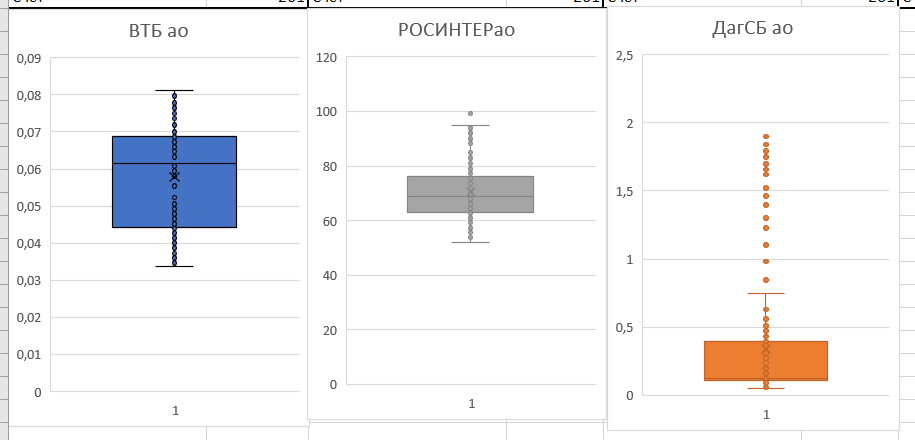
Более ясное понимание дает описательная статистика по логарифму доходности.



Далее логарифм объема торгов. Самый постоянный объем торгов у акций ВТБ, как и самый большой объем.

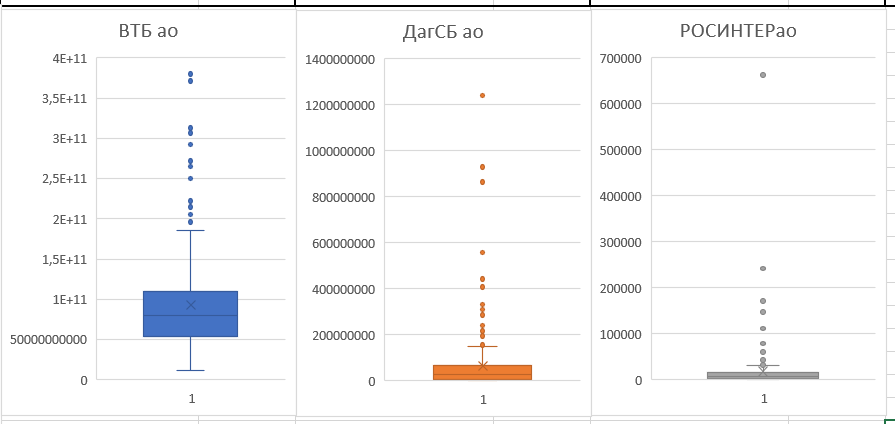


Диаграммы Ящик с усами для цен

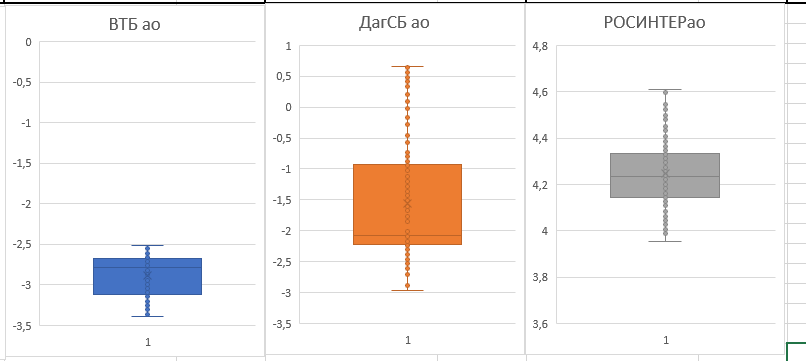


Как можно заметить, цены Акций ВТБ не имеют выбросов и распределены более равномерно. Росинтер имеет незначительное количество выбросов, а вот ДагСБ имеет целый период аномально больших цен – выбросов.

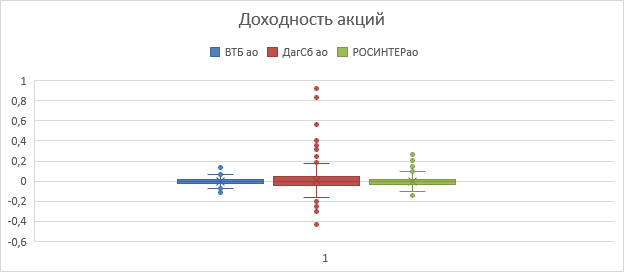
Из следующих диаграмм, построенных для обхемов торгов, видно, что все три компании имели аномально большие объемы торгов, тогда как аномально низких выбросов не наблюдается.



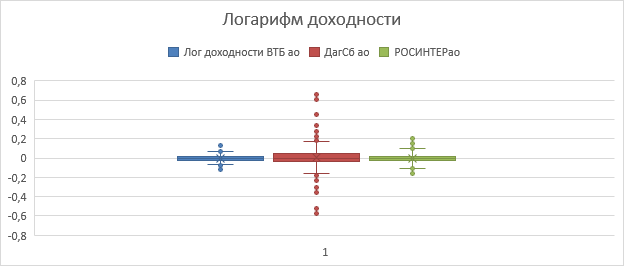
Далее логарифм цен



О жоходности акций можно сказать, что она почти у всех одинаковая, за исключением аномально больших выбросов в определенный промежуток времени у ДагСБ ао



То же самое можно видеть и для логдоходности



Диаграммы рассеяния для логарифма цен

И корреляции по ним

|  |  |  |  |
| --- | --- | --- | --- |
|  | *ВТБ ао* | *ДагСб ао* | *РОСИНТЕРао* |
| ВТБ ао | 1 |  |  |
| ДагСб ао | 0,620660633 | 1 |  |
| РОСИНТЕРао | 0,104895655 | -0,15194361 | 1 |

Здесь видим, что ВТБ ао и ДагСБ ао имеют более сильную зависимость между собой, чем в остальных случаях. При увеличении цен на акции ВТБ ао росли и цены на акции ДагСБ ал

Диаграммы рассеяния для логарифма объема торгов

|  |  |  |  |
| --- | --- | --- | --- |
| Логарифмы объемов торгов | | | |
|  | *ВТБ ао* | *ДагСб ао* | *РОСИНТЕРао* |
| ВТБ ао | 1 |  |  |
| ДагСб ао | 0,129967331 | 1 |  |
| РОСИНТЕРао | 0,12583388 | -0,019977193 | 1 |

Здесь сложно проследить какую-либо связь, характеристики друг от друга слабо зависят.

Для каждой характеристики построили гистограммы интервальных частот

Распределение цен РОСИНТЕРао больше всего похоже на нормальное.

Далее графики для объема

Данные диаграммы не отображают действительности вещей, потому что сильно искажены выбросами, ввиду чего имеем один интервал практически со всеми значениями, и интервалы с единичными аномальными значениями.

Логарифмированные цены уже более точно отражают картину, но все еще искажены выбросами, рисунки ниже

Далее логдоходности. Более всего на нормальное распределение похожи ВТБ и РОСИНТЕР, но диаграммы также не отражают реальное положение дел из-за выбросов.

Далее необходимо избавить от выбросов для Логдоходностей.

Для примера построим корреляционную матрицу признаков Логдоходности до удаления выбросов

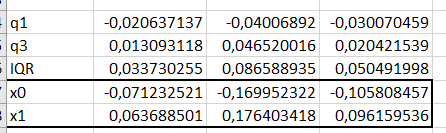
|  |  |  |  |
| --- | --- | --- | --- |
|  | *ВТБ ао* | *ДагСб ао* | *РОСИНТЕРао* |
| ВТБ ао | 1 |  |  |
| ДагСб ао | -0,006452271 | 1 |  |
| РОСИНТЕРао | 0,020434483 | 0,129605031 | 1 |

# Исследование логарифмических доходностей (логдоходностей) акций (удаление выбросов)

Далее проведем поиск и удаление выбросов.

Для каждой организации посчитаем

q1, q3 первый и третий Квартиль, найдем размах IQR и вычислим нижнюю границу x0 и верхнюю x1/



Далее все значения – строки, находящиеся вне этой границы, удалим.

Итого, после удаления, остается 199 строк (было удалено 62 значения).

Теперь проверим корреляцию после удаления выбросов

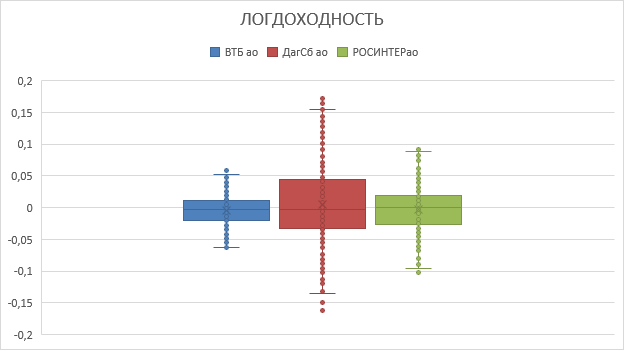
|  |  |  |  |
| --- | --- | --- | --- |
|  | *ВТБ ао* | *ДагСб ао* | *РОСИНТЕРао* |
| ВТБ ао | 1 |  |  |
| ДагСб ао | 0,050013471 | 1 |  |
| РОСИНТЕРао | 0,159906747 | 0,104673575 | 1 |

Теперь корреляция между признаками приняла более реальный вид.

Также можно увидеть разницу описательной статистики



Асимметрия и Эксцесс выборок уменьшились, также, как и стандартное отклонение, что говорит о более кучном распределении данных.



На диаграмме ящик с усами видно, что относительно очищенных данных появились новые, но уже не такие существенные выбросы, относительно которых мы уже можем дать более состоятельную оценку.

Далее посмотрим на гистограммы интервальных частот исправленных логдоходностей

Здесь уже можно явно заметить, что логдоходности без выбросов для компаний ВТБ ао и РОСИНТЕРао вероятно имеют нормальное распределение.

# Интервальные оценки параметров логдоходностей (в предположении нормального распределения), доверительные интервалы

Теперь исследуем доверительные интервалы логдоходностей трех акций.

Здесь мы видим, что нижние границы интервалов почти совпадают у всех трех акций, но у ДагСБ она немного ближе к 0, верхняя же граница у ДагСБ значительно выше, нежели у ВТБ и РОСИНТЕР. Следовательно инвестиции в акции ВТБ и РОСИНТЕР в худшем случае принесут одинаковые убытки, но акции ДагСБ чуть меньше убытков. А вот в лучше случае акции ДагСБ принесут гораздо больше прибыли, нежели акции двух других компаний. Причем акции ВТБ вообще не принесут прибыли, а только убытки.

Следовательно с 95% ной вероятностью можно утверждать, что акции ВТБ в среднем принесут только убытки, а акции ДагСБ принесут гораздо большую прибыль по сравнению с РОСИНТЕР.

Здесь мы можем видеть, что разброс логдоходности акций ВТБ минимален, в целом, то же самое можно сказать и о акциях РОСИНТЕР, хотя по сравнению с акциями ВТБ он все же больше, а вот разброс логдоходности ДагСБ ао существенно отличается от ВТБ и РОСИНТЕР. Следовательно ДагСБ ао имеет менее стабильное колебание акций и с большей вероятностью может отклониться от среднего значения.

# Проверка гипотезы о нормальности логдоходности по критерию Пирсона (хи-квадрат)

Для компании ВТБ ао:

H0: Нормальность присутствует

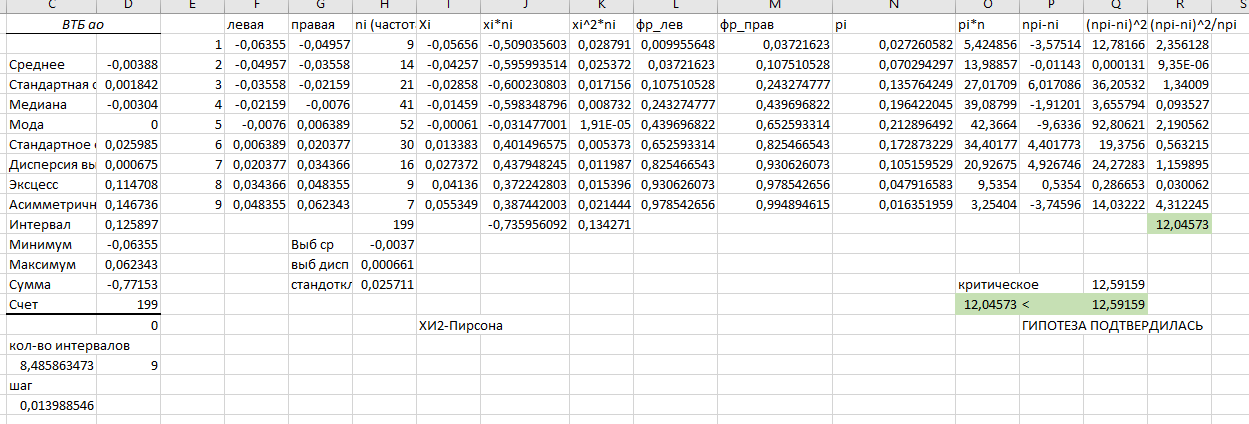
H1: распределение данных отлично от нормального

Альфа = 0,05

Сначала мы вычислили все необходимые параметры для выборки и разбили ее на 9 интервалов, посчитав для каждого частоту. Потом по методу моментов нашли необходимые параметры для нормального распределения, рассчитав выборочное среднее и выборочное стандартное отклонение, далее по формуле

Нашли сумму, то есть наблюдаемое значение 12,046.

Вычисляем критическую область для 95% вероятности по формуле ХИ2.ОБР(0,95;9-1-2), так как 9 интервалов и 2 параметра. Получаем, что наблюденной значение не попало в критическую область так как она правосторонняя, следовательно принимаем H0: нормальность присутствует, предпосылок для отклонения гипотезы нет.



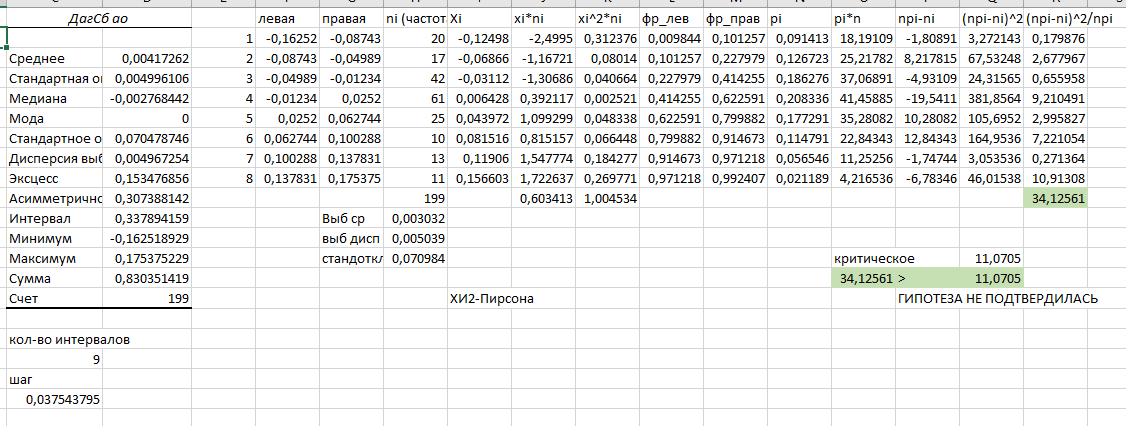
Аналогично предыдущему вычисляем наблюденное и критическое значение для компании ДагСБ ао.

Для компании ДагСБ ао:

H0: Нормальность присутствует

H1: распределение данных отлично от нормального

Альфа = 0,05



Здесь также должно было быть 9 интервалов, но частота первого была бы меньше 5, поэтому первый и второй интервал схлопываются в один, таким образом имеем 8 интервалов. Наблюденное значение 34 явно больше 11, сл-но попадает в критическую область и мы принимаем гипотезу H1: распределение данных отличается от нормального.

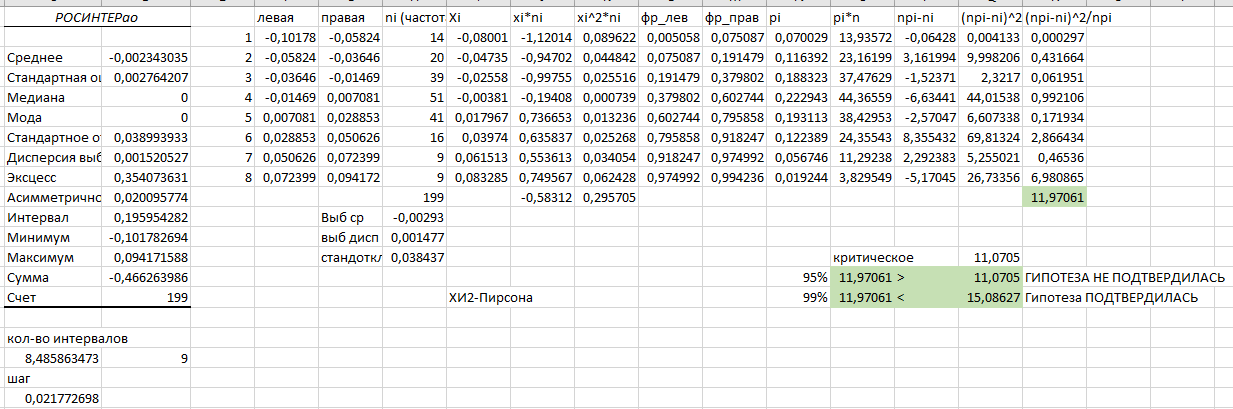
Следующая компания РОСИНТЕР ао.

Для компании РОСИНТЕР ао:

H0: Нормальность присутствует

H1: распределение данных отлично от нормального

Альфа = 0,05



Здесь аналогично с предыдущей ситуация, должно было быть бы 9 интервалов, но 1ый интервал содержал меньше 5 значений и слился со 2 интервалом в один. Рассчитанное значение 11,97 немного больше критического 11,075. То есть принимаем H1: распределение данных отлично от нормального с вероятностью 95%, но вот для 99% уровня наблюденное значение уже не попадает в критическую область, сл-но отвергнуть гипотезу с вероятностью 99% мы не можем.

# Проверка гипотезы о нормальности логдоходности по критерию Колмогорова-Смирнова

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Уровень значимости | 0,15 | 0,10 | 0,05 | 0,025 | 0,01 |
| Порядок *p*квантиля | 0,85 | 0,90 | 0,95 | 0,975 | 0,99 |
| Квантили порядка *p* для | 1,138 | 1,224 | 1,358 | 1,480 | 1,626 |
| Квантили порядка *p* для | 0,775 | 0,819 | 0,895 | 0,995 | 1,035 |

Для компании ВТБ ао:

H0: Нормальность присутствует

H1: распределение данных отлично от нормального

Альфа = 0,05

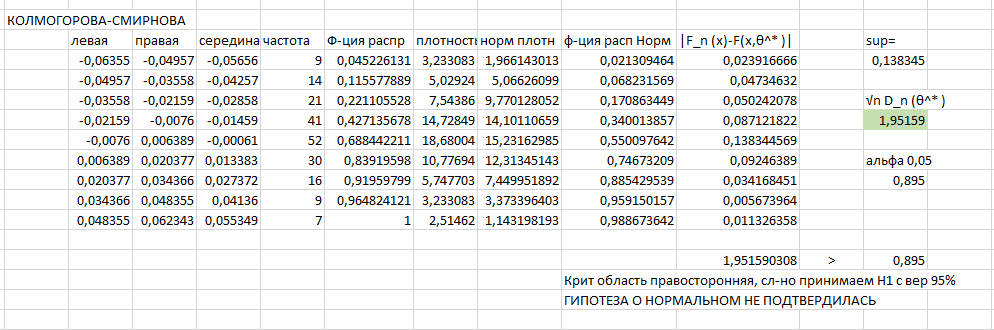
Берем уже вычисленные в предыдущем примере левую, правую границу, середину интервала и частоту. Далее по частоте находим интегральный процент и переводим в доли – столбец функция распределения.

Считаем плотность по формуле

Далее находим НОРМ.РАСП для плотности точечная, для функции распределения – интегральная.

Находим модуль разницы функции распределения эмпирической и функции распределения нормированной, и берем супремум из найденных значений и умножаем на корень объема выборки.

Найденное значение сравниваем с критическим, для 95% уровня критическое 0,895. Наблюдаемое 1,95, критическая область правосторонняя, следовательно принимаем гипотезу H1: распределение данных отлично от нормального.



Также построены диаграммы:

- гистограмма эмпирической плотности и график теоретической плотности

- гистограмма эмпирической ФР и график теоретической ФР

По ним можно заметить, что распределение близко к теоретическому нормальному.

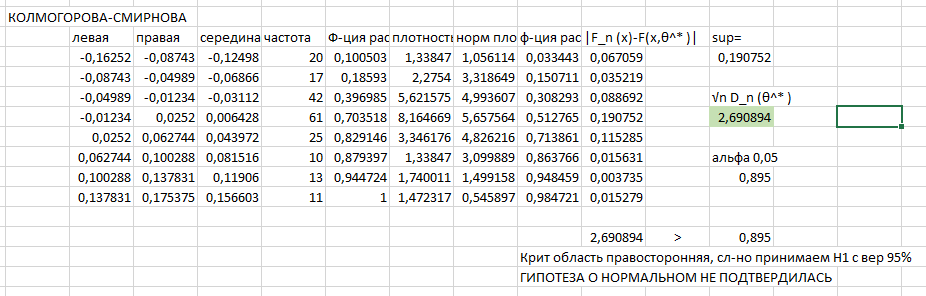
Аналогично предыдущему вычисляем наблюденное и критическое значение для компании ДагСБ ао.

Для компании ДагСБ ао:

H0: Нормальность присутствует

H1: распределение данных отлично от нормального

Альфа = 0,05



Здесь аналогично предыдущему принимаем H1: распределение данных отлично от нормального с вероятностью 95%.

Также построены диаграммы:

- гистограмма эмпирической плотности и график теоретической плотности

- гистограмма эмпирической ФР и график теоретической ФР

Здесь уже более заметно расхождение эмпирической функции и теоретической, больше отличается от нормального распределения.

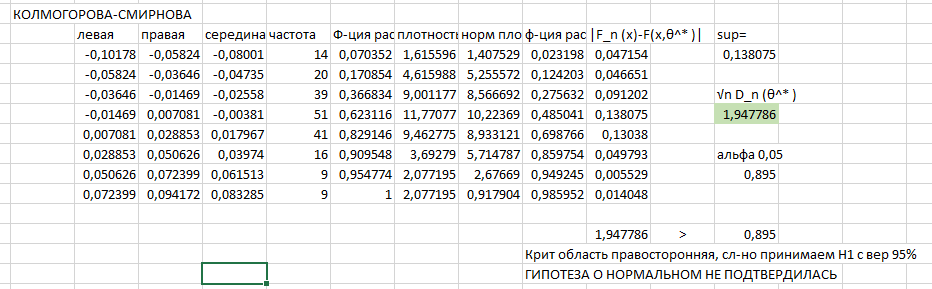
Аналогично предыдущему вычисляем наблюденное и критическое значение для компании РОСИНТЕРао.

Для компании РОСИНТЕР ао:

H0: Нормальность присутствует

H1: распределение данных отлично от нормального

Альфа = 0,05



Здесь аналогично предыдущему принимаем H1: распределение данных отлично от нормального с вероятностью 95%.

Также построены диаграммы:

- гистограмма эмпирической плотности и график теоретической плотности

- гистограмма эмпирической ФР и график теоретической ФР

Распределение близко к нормальному.

# Проверка гипотезы о равенстве нулю математического ожидания недельной логдоходности при альтернативное гипотезе о том, что оно больше нуля

Расчет теоретического значения производится по формуле

находится критическая точка распределения Стьюдентаt\_α (n-1), соответствующая уровню значимостиα и числу степеней свободы n-1. Нулевая гипотеза принимается, если

T≤t\_α (n-1)

и отклоняется при

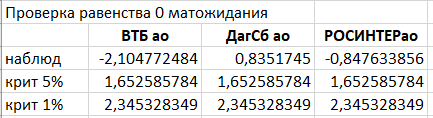
T>t\_α (n-1).

Гипотезы сформулированы следующим образом

**Н0 : m=0**

**Н1: m>0**

В результате проверки получаем следующую таблицу.



Для всех значений наблюденное меньше критического, следовательно везде принимаем H0: m=0.

Также данную проверку можно сделать при помощи Z.ТЕСТ



Все значения >0.1, следовательно как и было уже сказано везде принимаем H0.

# Проверка гипотезы о равенстве дисперсий за последний и предпоследний годы (двусторонний F-тест)

Гипотезы

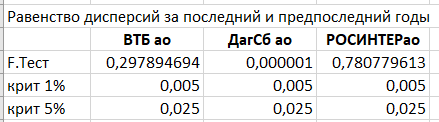
**Н0: дисперсии по годам равны**

**Н1: дисперсии по годам различаются**

В диапазонах F-теста указываем 2018 и 2019 год – логдоходность.

Он возвращает значение p-value, которое мы сравниваем с a/2. В результате сравнения получаем, что

p-value для ВТБ и РОСИНТЕР больше a/2, а вот для ДагСБ меньше a/2. Следовательно для 1 и 3 компании принимаем H;0 дисперсии равны по годам, а для ДагСБ H1: с вер 99% - дисперсии не равны по годам.



На основе предположений о равенстве дисперсий проверяем равенство средних.

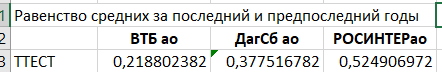
# Проверка гипотезы о равенстве средних значений логарифмической доходности за последний и предпоследний годы (двусторонний t-тест с информацией о равенстве дисперсий)

Гипотезы:

**Н0: средние по годам равны**

**Н1: средние по годам не равны**

Для 1 и 3 компании берем ТТЕСТ с типом 2 – дисперсии равны, а для ДагСб берем с типом 3 – дисперсии не равны на основании предыдущего пункта, получаем следующие p-value.



Все p-value > 0.1, поэтому принимаем H0: средние по годам равны. Во всех трех случаях.

# Сравнительный анализ логдоходностей трех компаний

Будет применен однофакторный дисперсионный анализ.

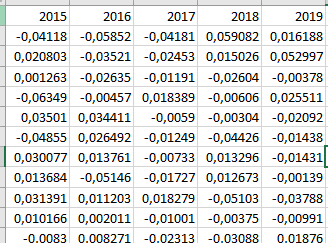
Для компании ВТБ ао:

Гипотезы:

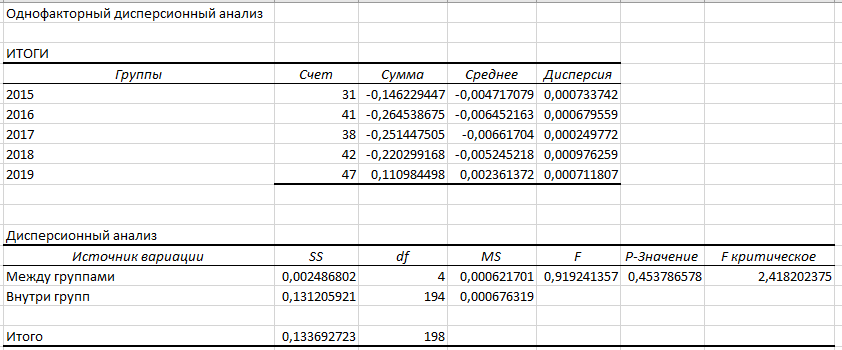
**Н0: m1=m2=m3=m4=m5**

**H1: средние не равны по годам**

Для начала разделим данные по годам



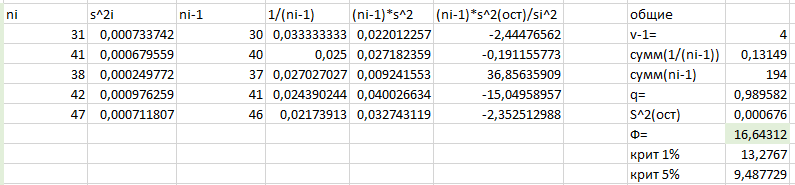
Затем проведем дисперсионный анализ



Для обоснованности выводов по дисперсионному анализу сначала необходимо рассчитать критерий Бартлетта.

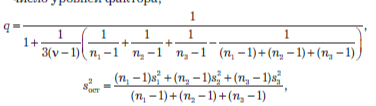
H0: дисперсии равны

H1: дисперсии не равны



По формуле рассчитываем Ф





Находим Критическое значение Ф через ХИ2.ОБР. Получаем, что наблюденное попадает в критическую область (правосторонняя). Следовательно H1, гипотеза о равенстве дисперсий не подтверждается и Применение дисперсионного анализа необоснованно.

Но если его все же применить, то p-value>0.1, сл-но принимаем H0 о равенстве средних, то есть логдоходность не зависит от года.

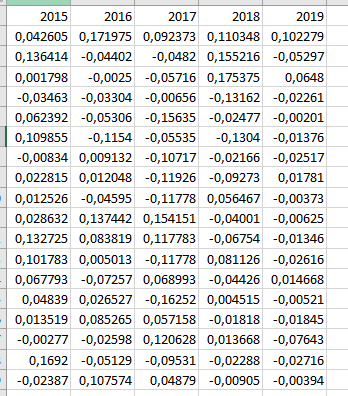
Аналогично проверим для ДагСБ оа:

Гипотезы:

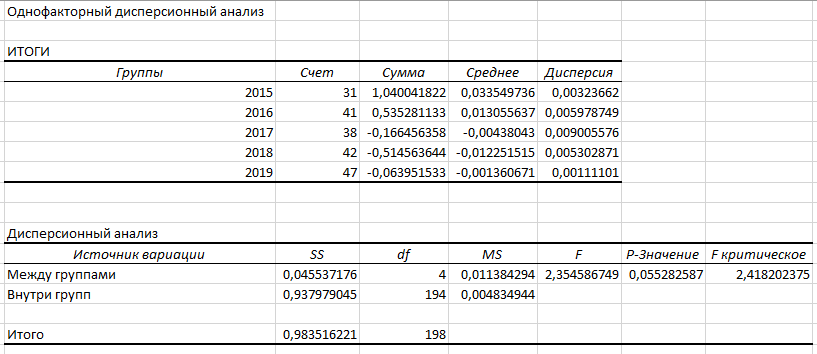
**Н0: m1=m2=m3=m4=m5**

**H1: средние не равны по годам**

Данные по годам



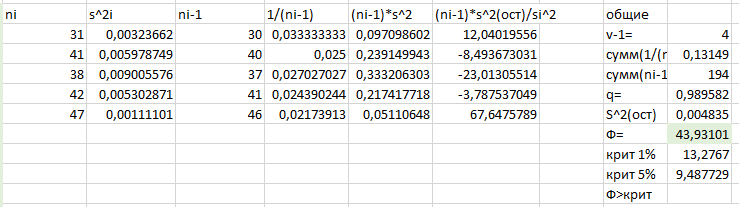
Дисперсионный анализ



Критерий Бартлетта.

H0: дисперсии равны

H1: дисперсии не равны



Снова применение дисперсионного анализа не обосновано, но если применить, то p-value между 0,1 и 0,05, то есть отклоняем H0 с вероятностью 90%, и логдоходность зависит от года.

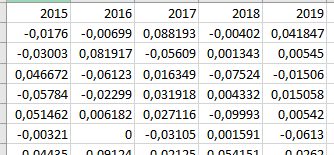
Аналогично проверим для РОСИНТЕРао:

Гипотезы:

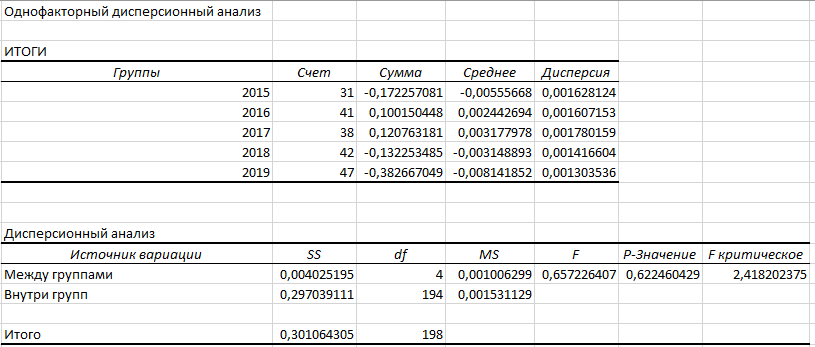
**Н0: m1=m2=m3=m4=m5**

**H1: средние не равны по годам**

Разделим данные



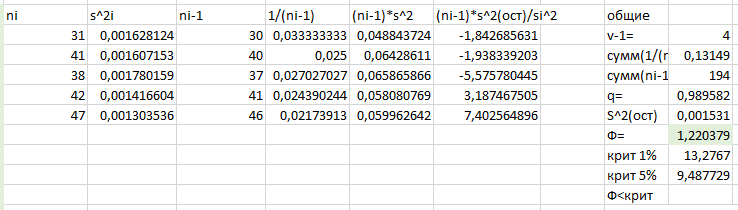
Далее дисперсионный анализ по этим данным



Определяем критерий Бартлетта

H0: дисперсии равны

H1: дисперсии не равны



В отличие от предыдущих компаний здесь наблюденное меньше критического, сл-но принимаем H0 и делаем вывод, что применение дисперсионного анализа обосновано.

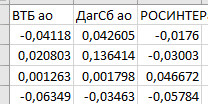
По p-value>0.1 делаем вывод, что средние равны, принимаем H0, сл-но логдоходность не зависит от года.

Далее дисперсионный анализ логдоходности в зависимости от названия

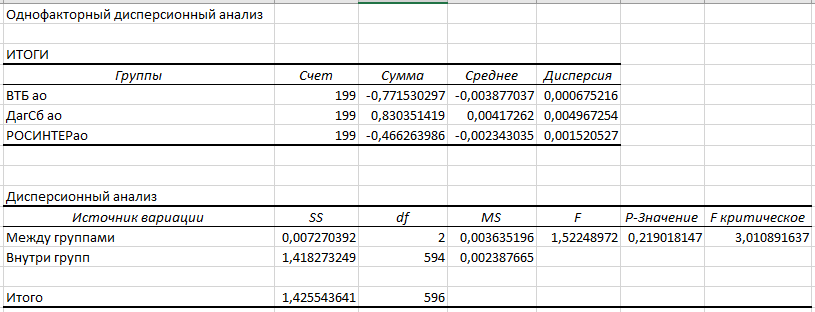
Возьмем данные логдоходности по трем компаниям

**Н0: m1=m2=m3**

**H1: средние не равны по компаниям**



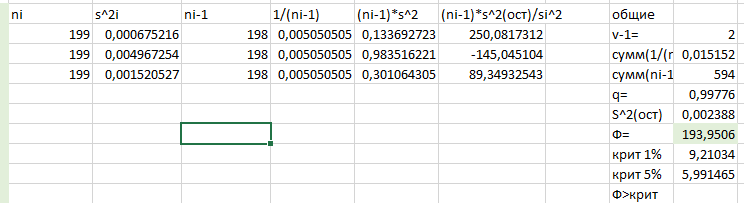
Выполним для них дисперсионный анализ



Далее рассчитаем критерий Бартлетта

H0: дисперсии равны

H1: дисперсии не равны



Здесь принимаем H1: дисперсии не равны, сл-но применение дисперсионного анализа не обосновано, но если его применить, то:

p-value > 0.1, принимаем H0, средние равны, а значит логдоходность не зависит от компании.

Эти различия нагляднее будут отображены на графике зависимости логдоходности от времени для каждой компании, они приведены ниже:

# Исследование тесноты связи между логдоходностями

Для этого исследования используем процедуру корреляции.

Выдвигается нулевая гипотеза

H0: r=0,

при альтернативной гипотезе

H1: r≠0.

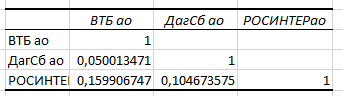
Далее вычисляем t

где

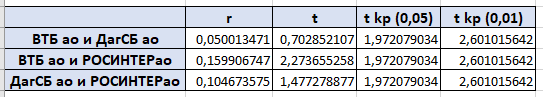
- распределение Стьюдента с n-2 степенями свободы.

Если |t|>t\_kp, то гипотеза H\_(0 ) отклоняется, в противном случае – не отклоняется.

Найдем матрицу корреляции с помощью Корреляция-Анализ Данных



Далее найдем t для каждого и t кр для альфа 0,01 и 0,05



Получаем, что для ВТБ и ДагСБ нет оснований отклонить гипотезу о наличии корреляции между логдоходностями компаний ни на уровне 0,05 ни на уровне 0,01

Для ВТБ и РОСИНТЕР есть основания отклонить гипотезу о наличии корреляции между логдоходностями на уровне значимости альфа =0,05

Для ДагСБ и РОСИНТЕР нет оснований отклонить гипотезу о наличии корреляции между логдоходностями ни на уровне значимости альфа = 0,05 ни на альфа =0,01