1. Выбрать из DataFrame 2 количественные и 2 качественные переменные.

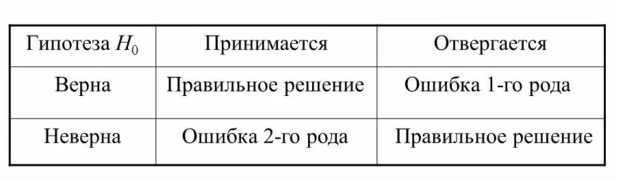
Проверить количественные переменные на нормальность.

Разбить значения количественных переменных относительно значений качественных.

Проверить полученные подвыборки на нормальность.

Реализовать для подвыборок t-test с помощью встроенной функции и вручную (реализовав формулу – тест стьюдента , ).

Задать вопрос: является ли та или иная величина норм распределенной или нет? Если да то с каким уровнем статистической значимости?



Как мы проверяем гипотезу?

1. Формулируем гипотезу

2. Определяем уровень значимости – вероятность ошибки первого рода

3. Когда мы проверяем какую-то гипотезу с помощью какого-то статистического критерия, этот критерий всегда имеет какую-то функцию, на распределение которой мы смотрим

Мы берем какую-то функцию от нашей выборки и смотрим на ее распределение

Для теста шапира-вилка например, берется функция W

Если ее значение попало куда-то в хвост распределения, мы отвергаем гипотезу

Типа надо нарисовать распределение статистики и ее значение

чем больше p-value тем лучше ?? (да)

p-value это площадь под графиком до значения w статистики

Для каждого теста его статистика распределена по какому-то закону

Мы не опровергаем, что величина распределена нормально. Мы это допускаем и принимаем на веру.

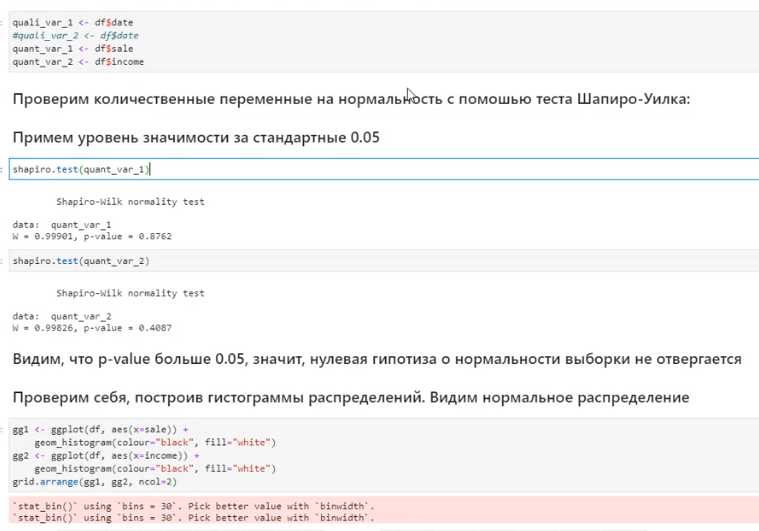
Вывести графики с распределениями переменных и статистики t-test.

Распределение фишера по ходу ?? хз

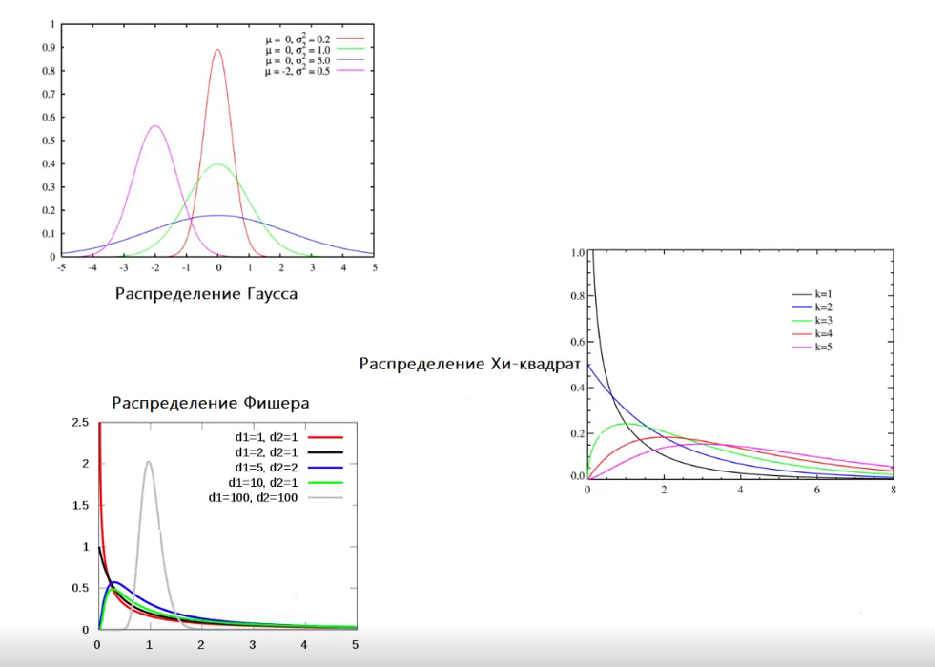
Гистограмму, плотность распределения

уровень статистической значимости показывает максимальную вероятность ошибки первого рода (??)

уровень статистической значимости это порог, при котором мы принимаем или отвергаем гипотезу



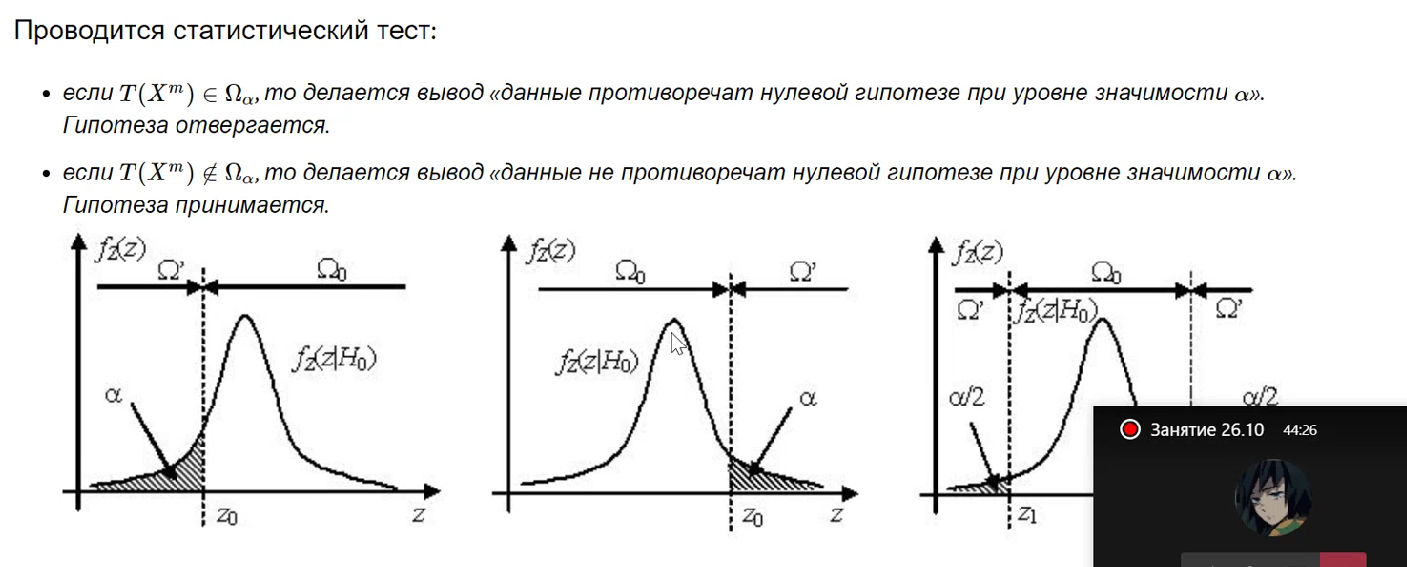
видимо такие графики



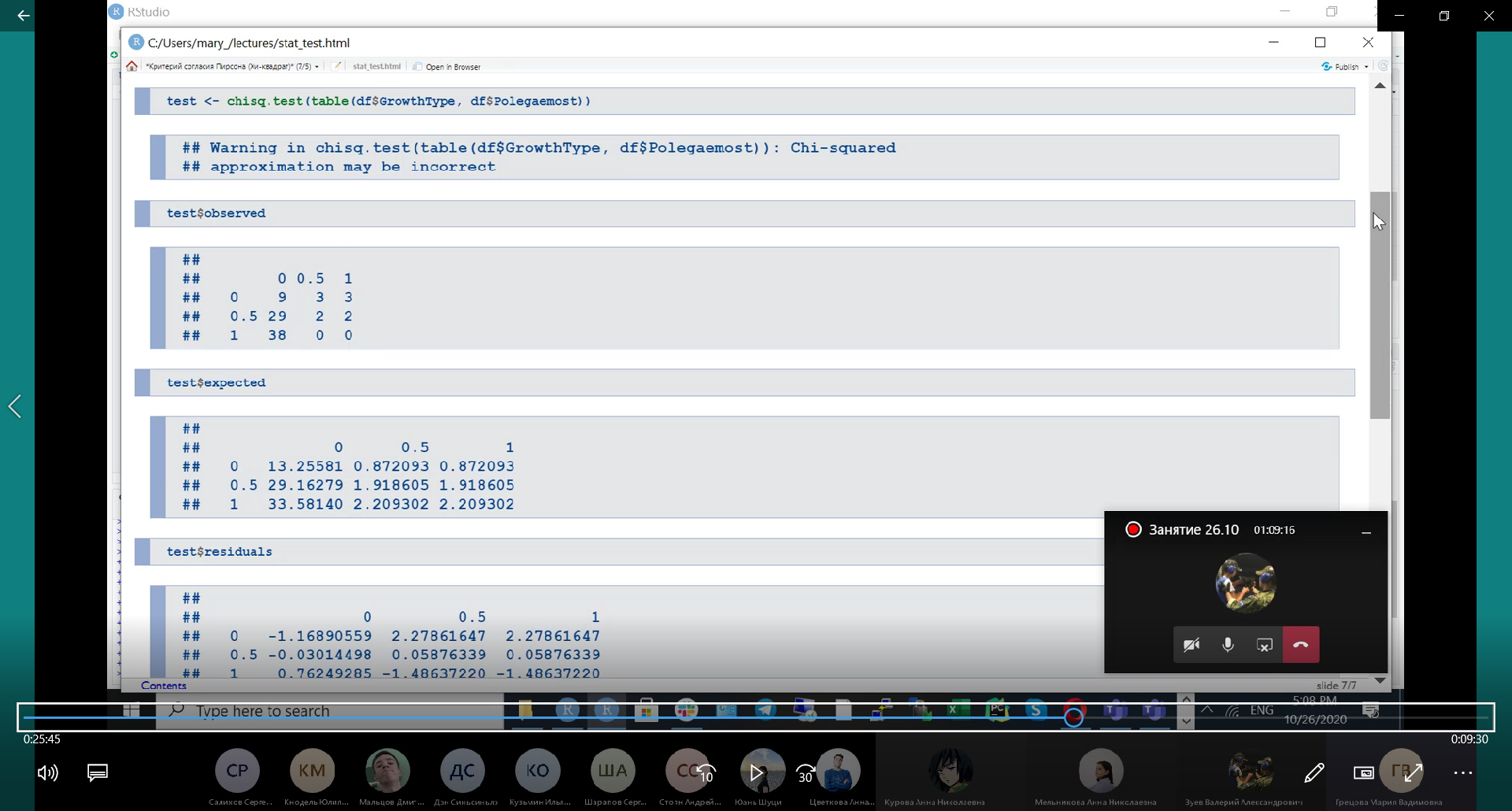
Чтобы сделать тест стьюдента, надо сначала провести тест шапира-вилко (shapiro.test()), потом тест фишера (проверка равенства дисперсий) var.test(): , H\_0 что числитель равен знаменателю.

sqrt

2. Выбрать 2 качественные переменные из старого DataFrame (??? какого??? по ходу салес), проанализировать зависимость между ними с помощью критерия Хи-квадрат (как у Баженова, лаба 7) 2мя способами – вручную и с помощью встроенной функции. chisq.test(table(…))



Сравнить результат с тем, который получился при использовании встроенной функции в R.

Построить для Вашего теста распределение статистики. На графике обозначить области принятия и отклонения гипотезы, уровень отсечения и значение статистики, полученное для Вашего статистического теста (см. рисунок из презентации).

Выводы из тестов – написать гипотезу, которую проверяем и вывод по цифрам; при каком уровне статистической значимости мы все проводили

Распределение хи-квадрат это распределение случайных величин имеющих нормальное распределение со средним 0 со стандартным отклонением 1

Для хи квадрат степень свободы это количество случ. величин, нормально распределенных, с помощью которых мы можем представить хи-квадрат распределение

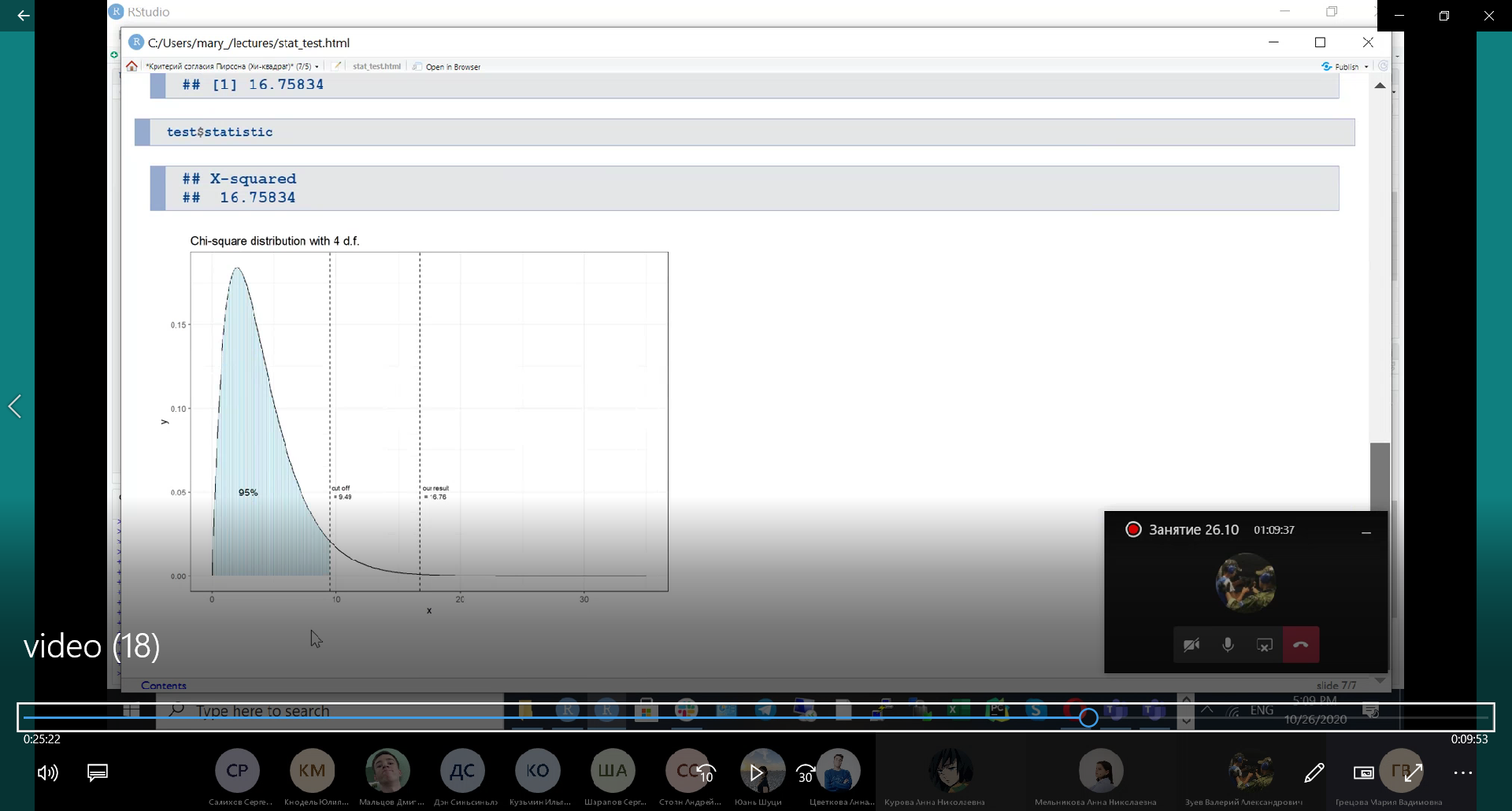
Фишер это два хи квадрата (их отношение), у него 2 степени свободы

Если две выборки:

df это степени свободы

Строим фишера для двух дф и гистограмму по выборке и саму функцию фишера

Тест Стьюдента – проверка гипотезы о равенстве средних



Пусть статистические ряды, построенные по двум сравниваемым выборкам, имеют вид, представленный в таблице:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Категория | 1 | 2 |  |  |  |
| выборка 1 |  |  |  |  |  |
| выборка 2 |  |  |  |  |  |
|  |  |  |  |  |  |

В критерии хи-квадрат мы разбиваем выборку на интервалы. В таблице категория обозначает номер интервала.

Через будем обозначать теоретические вероятности для каждой категории для гипотезы : «распределения совокупностей не различаются».

Составим статистику: **

Вероятности неизвестны, поэтому заменим их оценками, вычисленными при условии истинности гипотезы : *, где , .*

Тогда получим следующую статистику: **

Возьмем в качестве критической точки квантиль распределения Хи-квадрат уровня с степенью свободы. Тогда, если , то принимается гипотеза об однородности двух распределений, в противном случае различие распределений признается значимым.

Гистограмму, ящики с усами ???

плотности распределений

Когда передо мной стоит задача нарисовать распределение статистики, что мне рисовать?

типа если это тест фишера то это распределение фишера

зависит от статистического теста (надо просто фишера нарисовать короче)))0

нужно еще докинуть тест фишера туда ебать я не ебу откуда сука

статистика теста имеет распределение фишера

сначала нормальность, потом равенство дисперсий, потом только тест стьюдента t.test

<http://statistica.ru/theory/tablitsy-sopryazhennosti-vliyanie-privivki-na-kholernuyu-infektsiyu/>

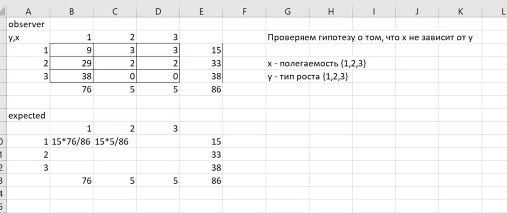
распределение статистики в t.test самим рассказать

Критерий согласия Пирсона проверяет подчиняется ли какое-то наблюдение какому-то закону

То есть ли зависимость или нет ли зависимости от ожидаемого

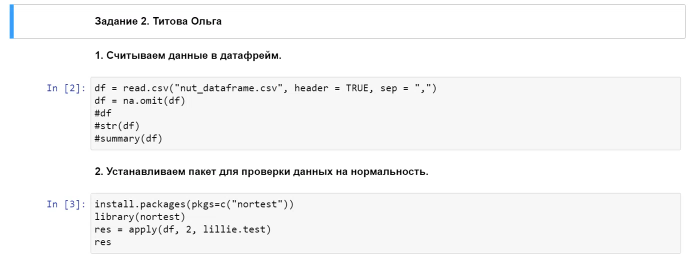
Первая таблица просто про те величины которые имеют две такие характеристики (их количество)

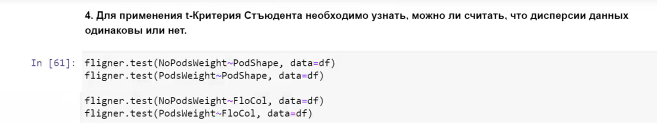
таблица сопряжений тупа

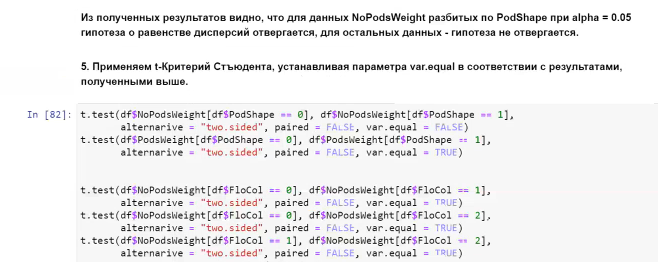


t.test – проверка равенства средних

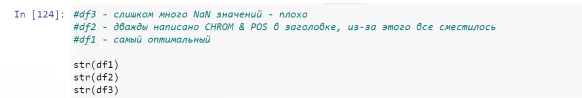
На одном графике и гистограмма, и плотность распределений

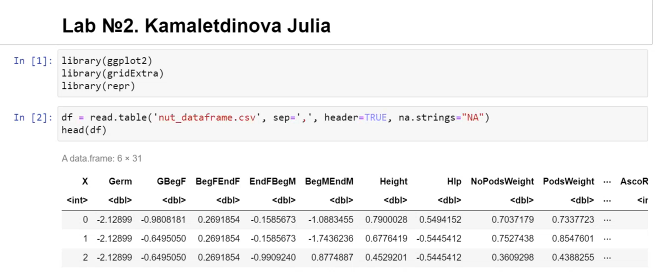


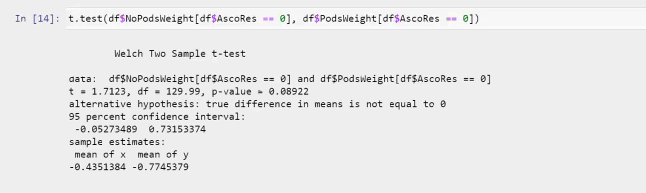


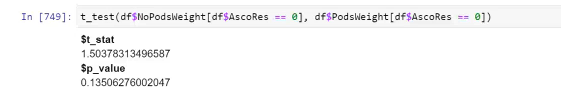


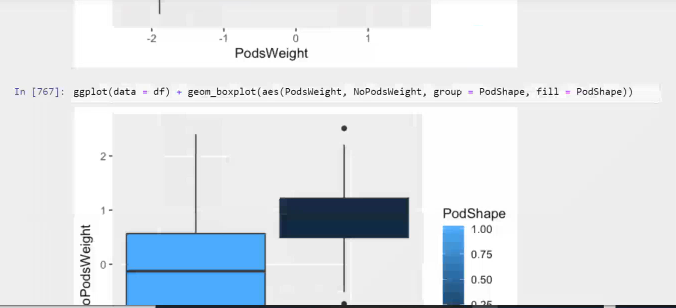




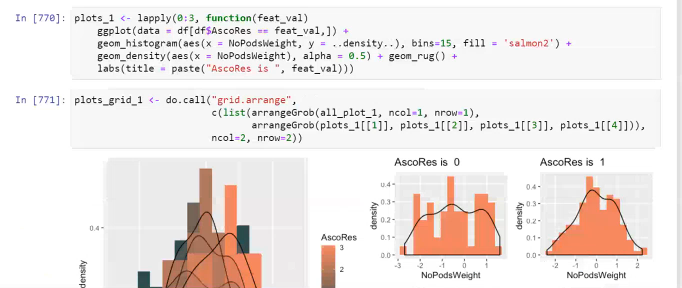


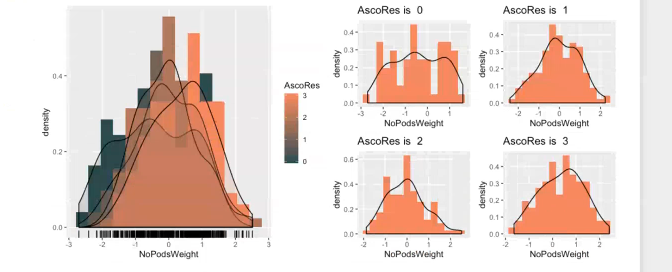












Для количественной переменной анализируем, является ли она нормальной или нет

зависит ли она от какой-то качественной переменной или нет

и то же самое для качественной с помощью теста хи-квадрат

t.test распределение статистики для него тоже нарисовать помимо фишера

Все датафреймы это nut\_dataframe.csv

!!!