«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Лабораторная работа №8 по теме: «Одновыборочный t-критерий»

Выполнил

Студент 1 курса

группы 09-115(3)

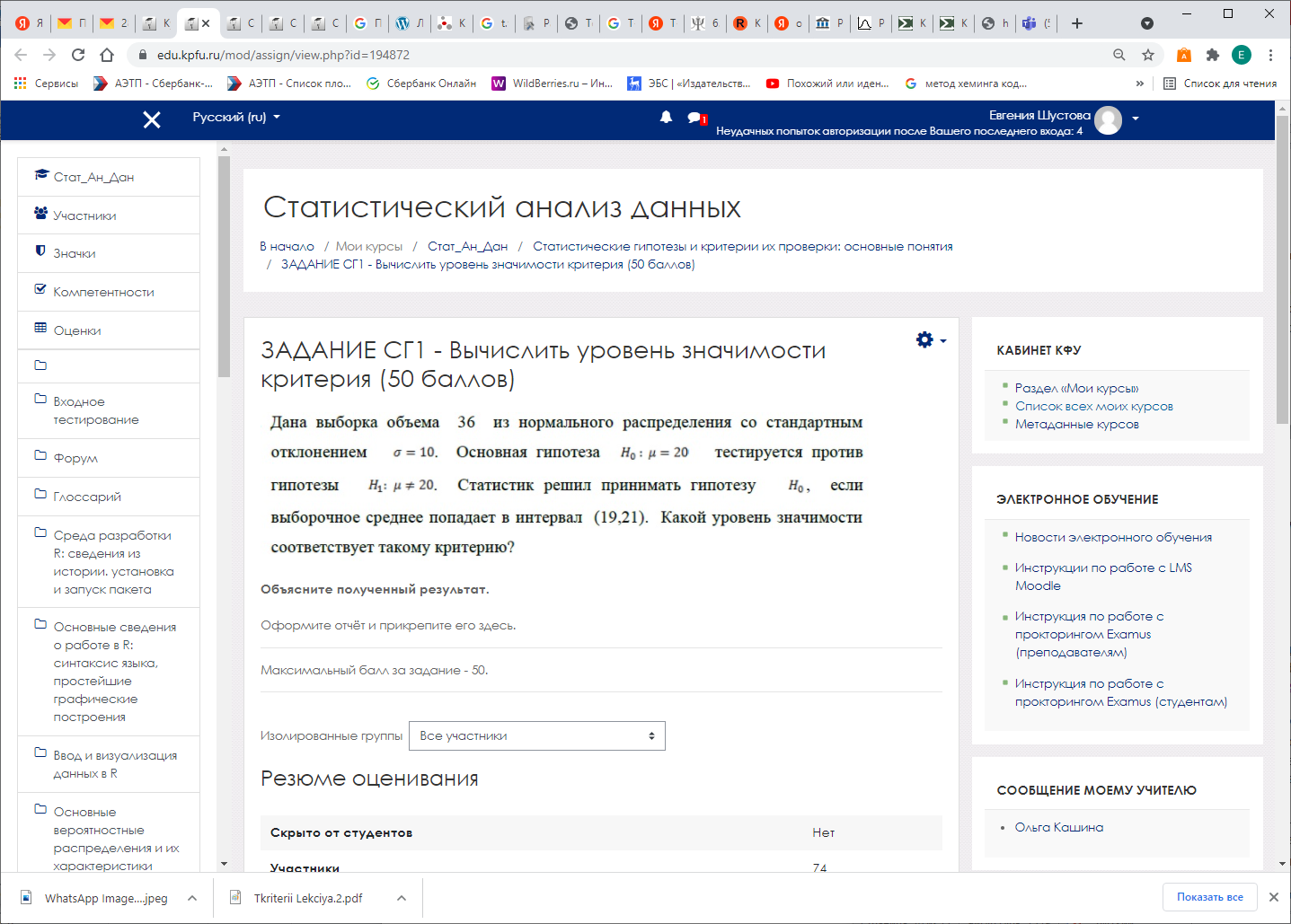
Зиновьев Е. А.

Преподаватель:

Шустова Е.П

Казань 2021

1. Выберите свои данные из репозитория данных, закон распределения которых является нормальным, приведите их краткое описание. Покажите, что выбранные Вами данные распределены по нормальному закону.
2. Сравните выборочное среднее этой выборки с заданным Вами значением (обозначим его через mu). Вычислите разность между выборочным средним значением и установленной нормой mu? Насколько различие статистически значимо с учетом уровня вариации значений в выборке?
3. Решите задачу (https://edu.kpfu.ru/mod/assign/view.php?id=194872):



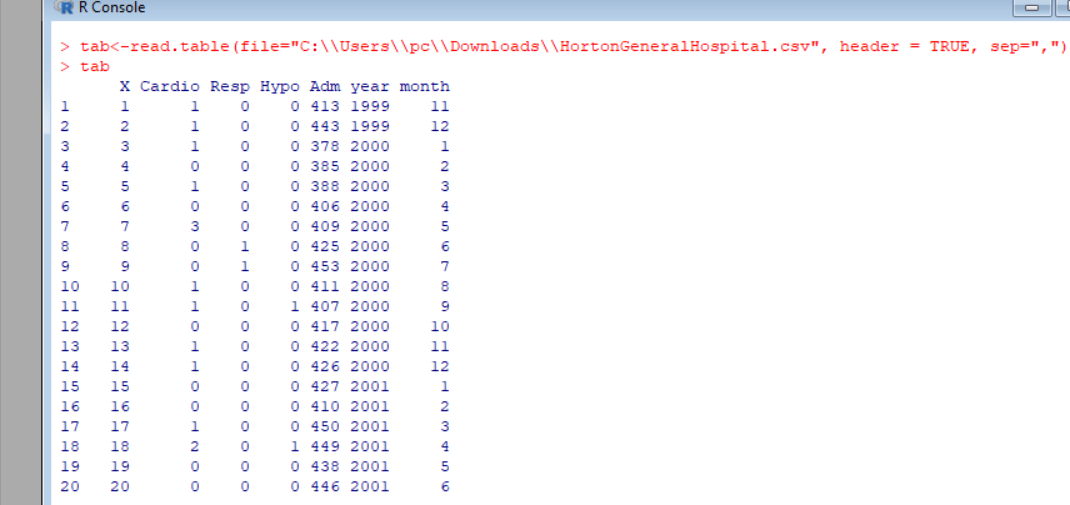
Ход работы

* + 1. Выбран набор данных:

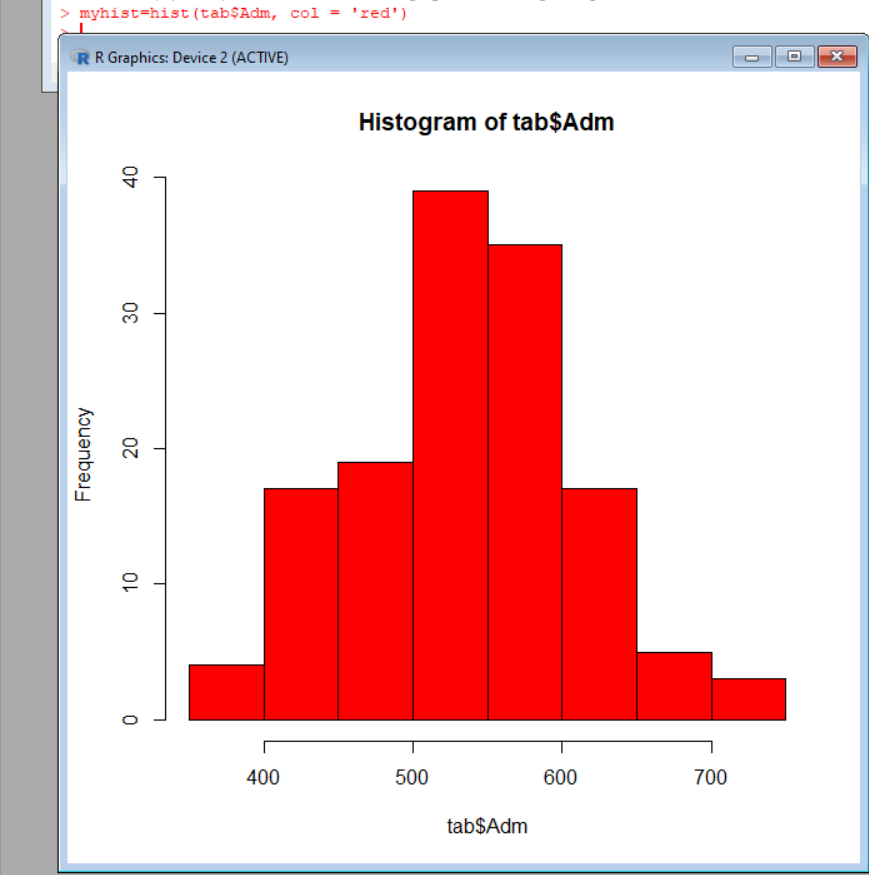
<http://archive.ics.uci.edu/ml/datasets/Horton+General+Hospital>

Cardio, Resp, Hypo - это ежемесячное количество переводов из отделения неотложной помощи в отделение реанимации, классифицированное в трех эксклюзивных категориях в больнице общего профиля Хортона: остановка сердечно-дыхательной системы, остановка дыхания, гипогликемическую остановку соответственно. Adm - это ежемесячное общее число пациентов, поступающих в отделение неотложной помощи.

Выведем этот датасет:

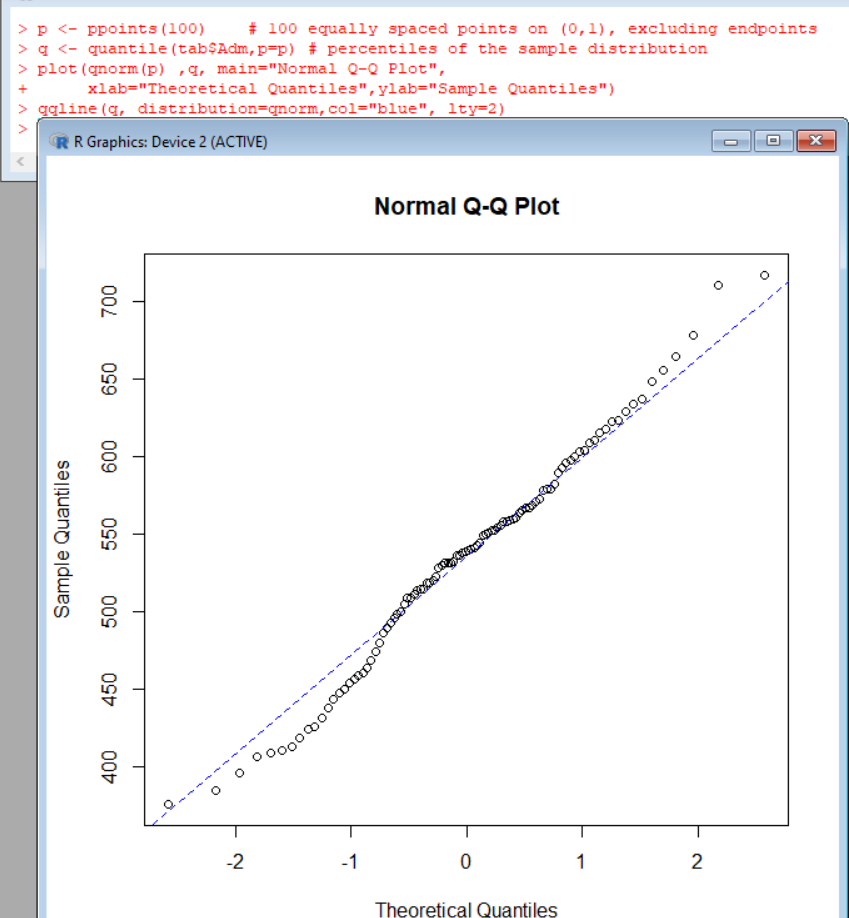


Для анализа распределения используем столбец Adm. Построим гистограмму распределения для этой переменной, чтобы графически проверить нормальность данных



По виду верхнего графика можно утверждать, что он имеет нормальное (из-за формы купола) распределение. При чем чуть более 70% значений лежит в промежутке между 500 и 600, в районе 1 стандартного отклонения.

Другим графическим методом проверки на нормальность является график q-q plot:



По графику также хорошо видно, что распределение является нормальным, несмотря на то, что нижняя часть нашей исследуемой лежит ниже теоретической нормальной кривой.

Существует целый ряд статистических тестов, специально разработанных для проверки нормальности распределения данных. В общем виде проверяемую при помощи этих тестов нулевую гипотезу можно сформулировать так: "анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение". Если получаемая при помощи того или иного теста вероятность ошибки р оказывается меньше некоторого заранее принятого уровня значимости (например, 0.05), нулевая гипотеза отклоняется.

В R реализованы практически все имеющиеся тесты на нормальность – либо в виде стандартных функций, либо в виде функций, входящих в состав подгружаемых пакетов. Примером базовой функции является shapiro.test(), при помощи которой можно выполнить широко используемый тест Шапиро-Уилка. Можно также перечислить функции из пакета nortest, реализующие другие распространенные тесты на нормальность:

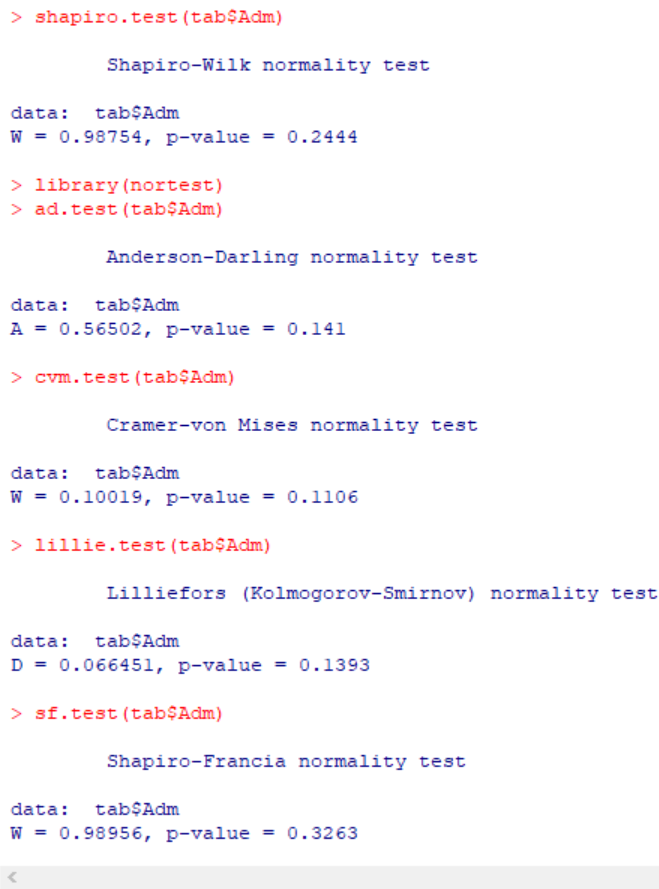
1. ad.test() - тест Андерсона-Дарлинга;

2. cvm.test() - тест Крамера фон Мизеса;

3. lillie.test() - тест Колмогорова-Смирнова в модификации Лиллиефорса;

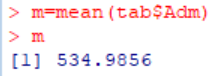
4. sf.test() - тест Шапиро-Франсия

Выполним серию тестов на проверку нормальности все той же выборки с использованием всех перечисленных критериев:

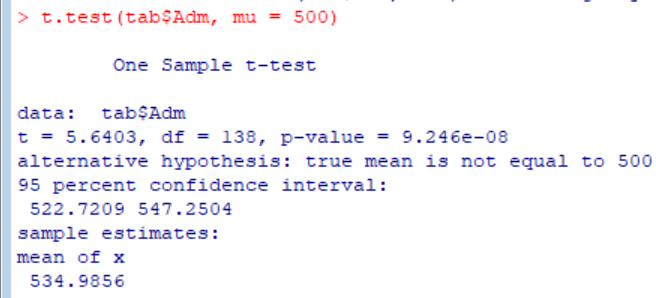


Как видим по p-value > 0.05, ни на одном тесте не была отклонена нулевая гипотеза о нормальности распределения данных.

1. Средняя ежемесячная норма количества пациентов неотложной помощи больницы Хортон примерно равно 500 человек, установленная для предупреждения и выявления массовых случаев серьезных заболеваний. Вычислим среднее количество пациентов для нашей выборки:

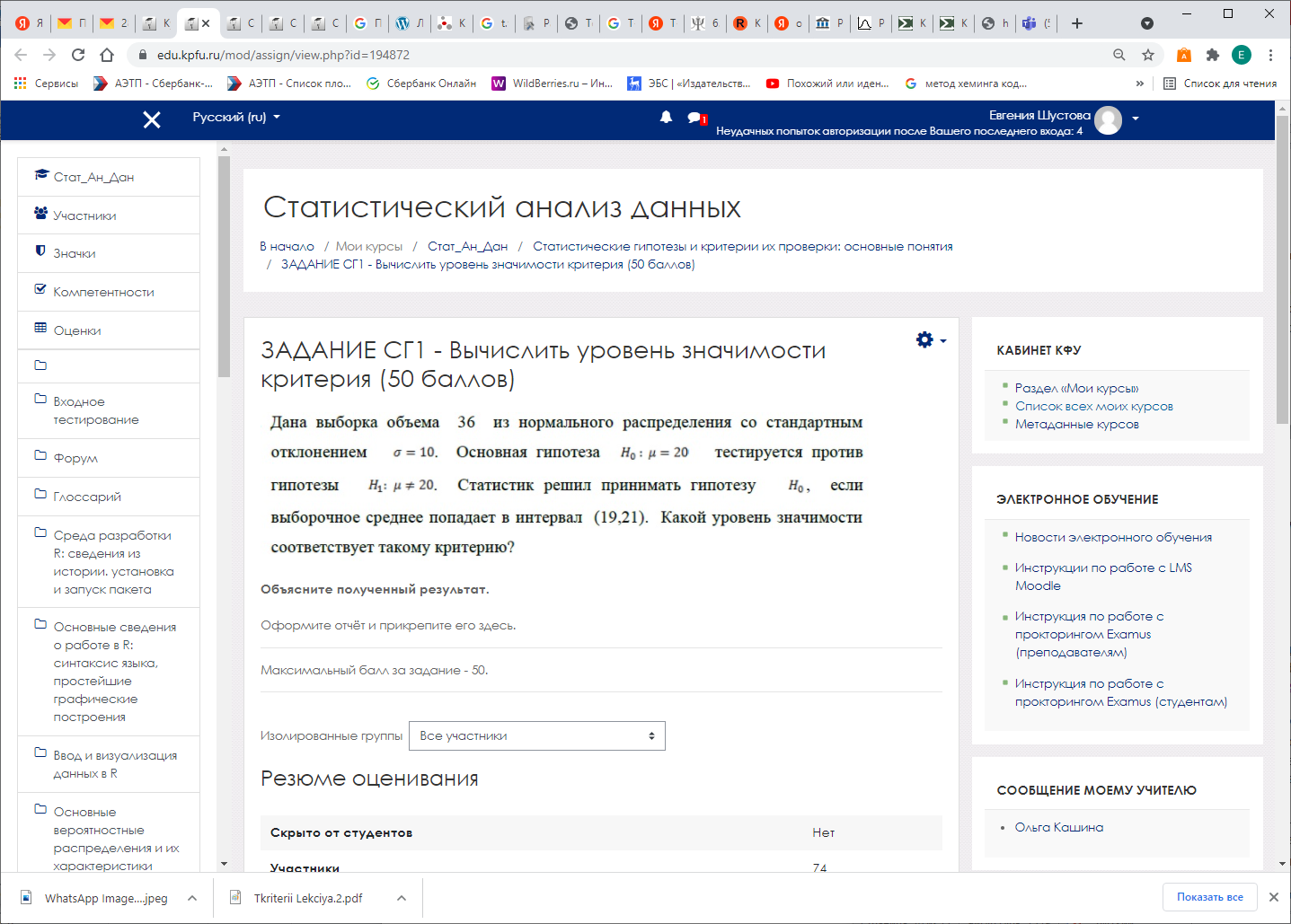


Зададимся вопросом: отличается ли это выборочное среднее значение от установленной нормы в 500 человек? Разница между нашим выборочным значением и этим нормативом составляет : 534.986 - 500 = 35. Но насколько эта разница статистически значима с учетом уровня вариации приведенных выше 139 значений (объема нашей выборки) ? Ответить на этот вопрос поможет одновыборочный t-тест. Как и другие варианты t-теста, одновыборочный тест Стьюдента выполняется в R при помощи функции t.test():

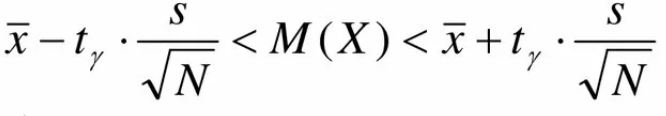


Видим, что для имеющихся выборочных данных t-критерий составляет 5.64 при 138 степенях свободы (df). Вероятность получить такое (либо большее) значение t при условии, что проверяемая нулевая гипотеза верна, оказалась весьма мала: p-value = 9.24 \*10-8 (во всяком случае, это меньше 5%). Следовательно, по правилу, представленному выше, мы можем отклонить проверяемую нулевую гипотезу о равенстве выборочного среднего значения нормативу и принять альтернативную гипотезу (alternative hypothesis: true mean is not equal to 500). Принимая это предположение, мы рискуем ошибиться с вероятностью менее 5%.

Помимо t-критерия, числа степеней свободы, р-значения и выборочного среднего (sample estimates: mean of x), программа рассчитала также 95%-ный доверительный интервал (95 percent confidence interval) для истинной разницы между выборочным средним значением ежемесячного количества пациентов неотложной помощи и установленным нормативом. Если бы мы повторили аналогичный тест много раз для 139 месяцев, то в 95% случаев эта разница оказалась бы в диапазоне от 522.72 до 547.25 человек.



Доверительный интервал находится как интервал от вычитания среднеквадратического отклонения выборки, помноженного на коэффициент Стьюдента и поделенного на корень из объема выборки, из среднего значения выборки до прибавления среднеквадратического отклонения выборки, помноженного на коэффициент Стьюдента и поделенного на корень из объема выборки, из среднего значения выборки:



То есть получим следующие неравенства:

20 – 10t/√36 < M < 20 + 10t/√36

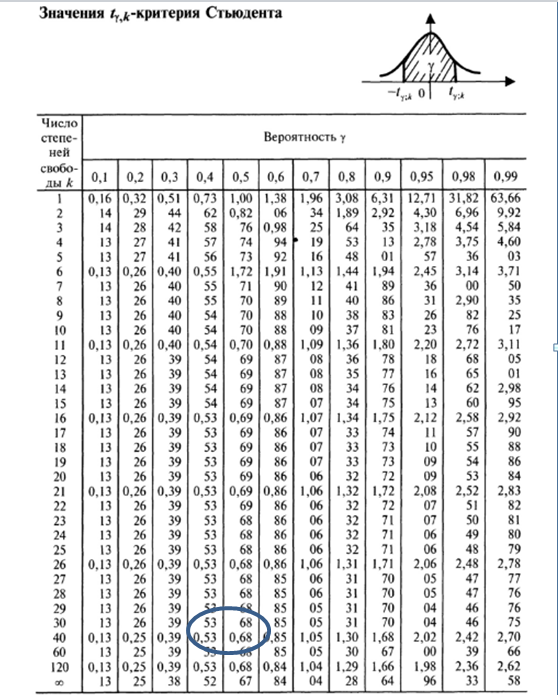
Для интервала (19, 21) надо приравнять:

20 – 10t/√36 = 19

20 + 10t/√36 = 21

Откуда t =0.6

По таблице коэффициентов t-Стьюдента найдем уровень доверия:



Число степеней для нашей выборки составляет 35 (36-1). Идем по строке 35 (у нас между 20 и 30), пока не увидим коэффициент близкий к 0.6 и поднимаемся наверх для уточнения степени доверия. В нашем случае он достигается между 0.4 и 0.5, где-то в районе 0.45.

Это значит, что нулевая гипотеза о среднем равном 20 при доверительном интервале (19, 21) будет отклоняться при степени доверия большему 45 %. Это значит, что чуть больше, чем в половине случаев может возникнуть ошибка.