

Задача 1. Какой регион (Country.Region) самый представленный?

Задача 2. Рассчитайте процент выздоровевших (Recovered) от суммы подтвержденных (Confirmed) и выздоровевших (Recovered). Результат запишите в виде нового столбца Percent.

Задача 3. Отберите (одной командой!) только такие наблюдения, где число выздоровевших (Recovered) больше числа подтвержденных (Confirmed), а число умерших (Deaths) равно 0. Сколько осталось строк?

Задача 4. Создайте список, в элементы которого положите все данные об одной стране (Country.Region). Таким образом, должен получиться список элементов которого будут представлять собой таблицы – один элемент списка – подвыборка исходного набора данных для одной страны. Проверьте, что число элементов списка равно числу уникальных значений из столбца Country.Region. Дайте каждому элементу листа имя вида: Country:номер_элемента листа

Задача 5. Создайте функцию, которая берет на вход таблицу (df) и число (A). Для таблицы df функция считает число строк. Далее полученное число строк делит на число A и умножает на 100. Примените полученную функцию к таблице, являющейся десятым элементом списка из задачи 4; число A – число строк в исходном наборе данных.

Задача 6. Примените функцию из задачи 5 к листу из задачи 4, где df – таблица из каждого элемента листа, а число A – число строк в исходном наборе данных, не прибегая к использованию циклов. Результаты запишите в другую таблицу, состоящую из 2 столбцов: имя элемента листа и результат функции.

Задача 7. Выберите все наблюдения за 09/21/2020, для которых число recovered больше 0 и отношение числа смертей к числу выздоровевших больше 0.1. Выведите три РЕГИОНА (Country.Region) с наибольшим числом смертей в отобранных наблюдениях.

Задача 8. Выберите 5 провинций Mainland China с наибольшим числом смертей за 07/21/2020. Проверьте на уровне значимости 0.05 гипотезу о том, что число выздоровевших этих провинциях распределено равномерно.

Задача 9. Проверьте гипотезу о том, что за 07/21/2020 число умерших в среднем в провинции Mainland China меньше, чем число умерших в среднем в штате США (US). Нет оснований предполагать равенство дисперсий. Можно использовать тест Стьюдента даже если правильнее было бы использовать нормальный тест.

Задача 10. 1) Напишите функцию genSample(N, mu, sd), которая генерирует N выборок размера 100 из нормального распределения со средним mu по-умолчанию равным 10 и стандартным отклонением sd по умолчанию равным 5. 2) Запустите функцию для N=10000. Для каждой выборки проверьте с помощью теста Стьюдента гипотезу о равенстве среднего 0. Выведите 5 наименьших полученных p-value

Задача 11*. Для каждой СТРАНЫ (Country.Region), где за 08/21/2020 число confirmed больше 1000 и Deaths больше 100 проверьте гипотезу о том, что доля смертей за 08/21/2020 не больше 0.1. Выведите страну, для которой p-value получился минимальным. Если таких стран несколько, выведите ту из них, где больше всего Confirmed случаев за этот день. Указание: вам поможет функция pbinom. Да, ее можно применять в mutate

Задача 12*. На одном из этапов развития философии науки, философы высказали идею, что цикл жизни любой теории выглядит так:
теория появляется
накапливаются доказательства в ее пользу
потом встречается противоречащий теории факт, и мы эту теорию полностью отвергаем

Допустим, вы хотите проверить, что ваше лекарство работает и понижает некий показатель X.

Допустим, что лекарство понижает X на 10.

Вы ставите эксперимент с $\alpha=0.05$, дисперсию X ваше лекарство не меняет, и она у измерений равна 7. Измерения распределены нормально.
Всего вы берете 20 пациентов (наблюдаете их до и после).
Какова вероятность ошибочно не отвергнуть H_0 ?

Теперь другой экспериментатор полностью повторяет ваш эксперимент. Какова вероятность того, что и он, и вы (верно) отвергнете H_0 .

Ответ можете вывести теоретически и подсчитать функцией в R, а можете получить при помощи корректной симуляции в R