Задача 4.1

Северилов Павел, 674

25 марта 2020 г.

Рассмотрим данные по числу заболевших и выздоровевших от короновируса в разных странах. Требуется проверить гипотезу о том, что число выздоровевших людей в странах не зависит от числа заболевших в стране.

1 Решение

Проверка зависимости признаков означает, что надо пользоваться корреляциями. Выборка достаточно небольшая ($n = 26$). Логично предположить, что в нашем случае зависимости скорее всего не линейны, поэтому корреляция Пирсона не подходит. Поэтому будем использовать корреляцию Кендалла, как лучше подходящую для поиска нелинейных взаимосвязей, чем корреляция Спирмена.

- Постановка задачи:
  $X_1^n = (X_{11}, \ldots, X_{1n})$ – заболевшие
  $X_2^n = (X_{21}, \ldots, X_{2n})$ – выздоровевшие
  $H_0 : \tau_{X_1X_2} = 0$ – нет взаимосвязи
  $H_1 : \tau_{X_1X_2} <\neq> 0$

- Статистика $T\left(X_1^n, X_2^n\right) = \hat{\tau}_{X_1X_2} = 1 - \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^{n} \left[[X_{1i} < X_{1j}] \neq [X_{2i} < X_{2j}]\right] = 1 - \frac{2}{325} \sum_{i=1}^{25} \sum_{j=1}^{26} \left[[X_{1i} < X_{1j}] \neq [X_{2i} < X_{2j}]\right]$ – выборочный коэффициент корреляции Кендалла

- При $n > 10$ нулевое распределение аппроксимируется нормальным распределением $\mathcal{N}\left(0, \frac{2(2N+5)}{9N(N-1)}\right)$. Наши данные удовлетворяют условию на n, поэтому мы можем приближенно так записать нулевое распределение
$$\mathcal{N}\left(0, \frac{57}{2925}\right)$$

- Правило принятия решения при уровне значимости $\alpha = 0.05$: нулевая гипотеза отвергается, если $|\hat{\tau}| \geq \Phi^{1-\alpha/2}_{\mathcal{N}\left(0, \frac{57}{2925}\right)}$, где $\Phi^{1-\alpha/2}$ – $(1 - \alpha/2)$-квантиль. Или что то же самое:
$$p(\hat{\tau}) = 2\left(1 - F_{\mathcal{N}\left(0, \frac{57}{2925}\right)}(|\hat{\tau}|)\right) \leq 0.05$$

- Проверим гипотезу с помощью scipy.stats.kendalltau. Результат: correlation = 0.2835, pvalue = 0.0443. Видим, что pvalue $< \alpha = 0.05$, т.е. при заданном уровне значимости α гипотеза отвергается.

  Т.к. данные по Китаю сильно отличаются от остальных – выглядит, как выброс, то посмотрим на результат работы теста на данных без Китая: correlation = 0.223, pvalue = 0.122. Сам

кэффициент корреляции не сильно изменился, а вот pvalue значительно – теперь нельзя отвергнуть гипотезу на уровне $\alpha = 0.05$.

- Найдем на уровне значимости $\alpha = 0.05$ зависимость мощности критерия от истинного значения статистики. Для этого будем генерировать такие наборы данных X_1, X_2, чтобы получить всевозможные значения $\hat{\tau}$ ($\tau \in [-1, 1]$). Будем считать корреляцию для сэмплов из распределений $X_1 \sim \mathcal{N}(0, 1), X_2 \sim \pm X_1 + \alpha \cdot \mathcal{N}(0, 1)$, где будем варьировать α. Далее проводим по 1000 испытаний для каждого варианта τ (т.е. для каждого α) и считаем мощность.

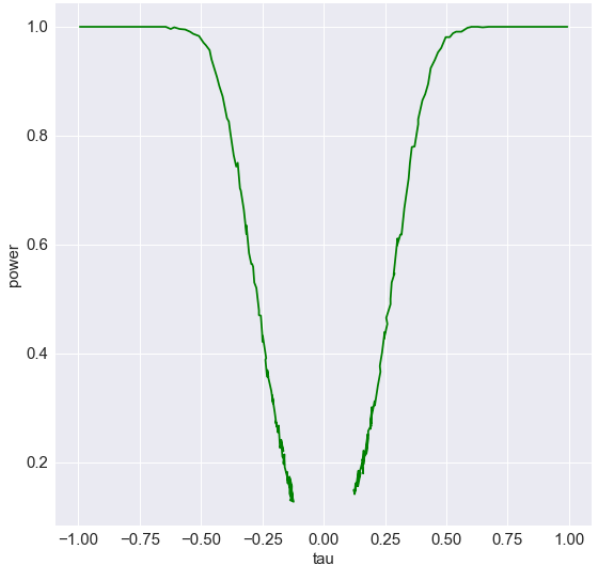

Рис. 1: Зависимость мощности критерия от истинного значения статистики

2 Вывод

Отвергли гипотезу о том, что между числом выздоровевших людей в странах и числом заболевших нет взаимосвязи – данные ей противоречат. Это значит, что между данными группами людей есть монотонная зависимость. Корреляция положительна, поэтому вероятнее всего, верно, что чем больше заболевших, тем больше выздоровевших. Большую роль в решении сыграли данные по Китаю – без него нельзя отклонить гипотезу H_0. Скорее всего, это связано с тем, что значения в Китае сильно отличаются от остальных и более явно показывают тенденцию рост заболеваемых→ рост выздоровевших, хоть ситуация в Китае уже на другом этапе в отличие от других рассматриваемых стран.

По построенной зависимости мощности от τ видим, что критерий имеет мощность 1 при абсолютном значении корреляции ≥ 0.5, т.е. в этих случаях критерий почти всегда отвергнет гипотезу о том, что нет зависимости между признаками, если она все-таки есть. Но при малых значениях коэффициента корреляции, похожих на полученные в задаче, критерий вероятнее всего не отклонит гипотезу, что нет взаимосвязи между признаками, когда эта взаимосвязь на самом деле есть. Этому соответствует случай, когда мы убрали данные о Китае: гипотезу H_0 мы не можем отклонить, но монотонная зависимость присутствует.