Задача 4.1

Рассмотрим данные из табллицы 1 по числу заболевших и выздоровевших от короновируса в разных странах. Требуется проверить гипотезу о тому, что число выздоровевших людей в странах не зависит от числа заболевших в стране.

Требуется:

- записать задачу формально;
- предложить статистику для решения данной задачи;
- записать приближенно нулевое распределение данной статистики;
- записать явно правило принятия решения на основе статистики и нулевого распределения для обеспечения уровня значимости $\alpha = 0.05$;
- проверить гипотезу по записанному критерию, для данных из условия. Противоречат ли они гипотезе?

На уровне значимости $\alpha = 0.05$ найти зависимость мощности критерия в зависимости от истинного значения статистики.

Решение

Требуется проверить независимость числа выздоровевших от числа заболевших при малом количестве данных ($n = 26$). При таком условии лучше подходит коэффициент корреляции Кендалдла (она точнее оценивается).

Постановка

- $X_1^n = (X_{11}, \dots, X_{1n})$ - заболевшие
- $X_2^n = (X_{21}, \dots, X_n)$ - выздоровевшие
- $H_0 : \tau_{X_1 X_2} = 0$
- $H_1 : \tau_* X_1 X_2 \neq 0$

Статистика

$\hat{\tau}_{X_1 X_2} = 1 - \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^{n} [[X_{1i} < X_{1j}] \neq [X_{2i} < X_{2j}]] = \frac{C-D}{C+D}$

- C - число согласованных пар,
- D - число несогласованных пар

При $n > 10$ ее можно аппроксимировать нормальным распределением $N\left(0, \frac{2(2n+5)}{9n(n-1)}\right) \sim N\left(0, \frac{57}{2925}\right)$

Решающее правило

Отвергнуть нулевую гипотезу, если

$|\hat{\tau}| \geq \bar{\Phi}^{1-\alpha/2}$, где$\bar{\Phi}^{1-\alpha/2} - \left(1 - \frac{\alpha}{2}\right) - \text{квантиль распределения } \mathcal{N}\left(0, \frac{57}{2925}\right)$

Тогда

$p(\hat{\tau}) = 2\left(1 - F(|\hat{\tau}|\right) \leq \alpha$

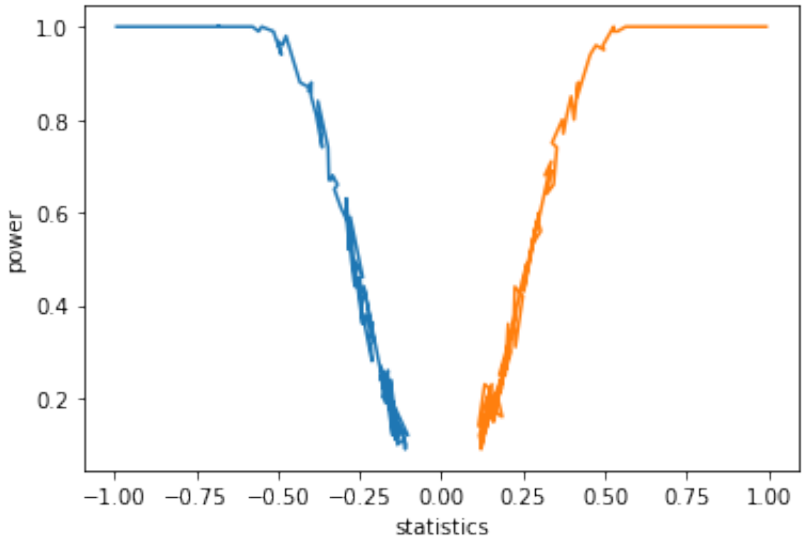
```
from scipy import stats as st
res = st.kendalltau(data['заболевшие'].values, data['выздоровевшие'].values)
print(res)
alpha = 0.05
if res.pvalue <= 0.05:
  print('Гипотеза о независимости отвергается')
else:
  print('Гипотеза принимается')

>>> KendalltauResult(correlation=0.28351110894619114,
pvalue=0.044325607642096566)
>>> Гипотеза о независимости отвергается
```

Учитывая, что Китай отличается от других стран большим опытом борьбы с инфекцией, интересно посмотреть на данные без Китая.

```
without_china = data.drop(labels=[5])
res = st.kendalltau(without_china['заболевшие'].values,
without_china['выздоровевшие'].values)
print(res)
alpha = 0.05
if res.pvalue <= alpha:
  print('Гипотеза о независимости отвергается')
else:
  print('Гипотеза принимается')

>>> KendalltauResult(correlation=0.22299333487183037,
pvalue=0.12212105747986692)
>>> Гипотеза принимается
```

```
## Мощность как функция от статистики
from matplotlib import pyplot as plt
num_exp = 100

for coef in [-1, 1]:
  t = []
```

```
  p = []
  for noise in np.linspace(0.01, 5, 100):
    tau = []
    rejected = 0
    for _ in range(num_exp):
      X1 = np.random.randn(26)
      X2 = coef*X1 + noise*np.random.randn(26)

      res = st.kendalltau(X1, X2)
      tau.append(res[0])
      if res[1] <= alpha:
        rejected += 1
    t.append(np.mean(tau))
    p.append(rejected/num_exp)
  plt.plot(t, p)

plt.xlabel('statistics')
plt.ylabel('power')
plt.show()

>>>
```



Вывод

1. В Китае, больше выздоровевших, потому что там больше заболевших, про остальные страны такое говорить рановато.
2. Про критерий: когда коэффициент корреляции по модулю стремится к единице, критерий наверняка отклонит нулевую гипотезу о независимости выборок при ее наличии. Но при малых значениях статистики, критерий, наоборот, с большей вероятностью нулевую гипотезу примет даже когда зависимость между признаками есть, что и произошло при рассмотрении стран без Китая.