Фамилия и имя: Вареник Наталия

Прикладной статистический анализ данных

Лабораторная работа 1

Задача 4.1 Имеются данные 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. Переформулируем независимость через коэффициент корреляции. Поскольку коэффициент корреляции является статистической мерой зависимости случайных величин, то проверяемую гипотезу можно переписать в следующем виде:

$$H_0 : \rho_{X_1X_2} = 0$$
$$H_1 : \rho_{X_1X_2} \neq 0$$

Здесь X_1, X_2 – это выборки, отвечающие за число заболевших и выздоровевших в стране соответственно, они связанные, ρ – коэффициент корреляции, в качестве которого будем использовать корреляцию Спирмена, поскольку он не требует нормальности данных. Для проверки описанной гипотезы воспользуемся критерием Стьюдента. Для этого рассмотрим статистику:

$$T = \frac{\rho_{X_1X_2}\sqrt{n-2}}{\sqrt{1-\rho^2_{X_1X_2}}}$$

Эта статистика при $n \to \infty$ в условии истинности нулевой гипотезы имеет распределение $St(n-2)$. Зададим крическую область, соответствующую уровню значимости $\alpha = 0.05$, при попадании реализации статистики в которую нулевая гипотеза отклоняется:

$$U_\alpha = (-\infty, t_{\frac{\alpha}{2}}) \cup (-t_{\frac{\alpha}{2}}, +\infty),$$

где $t_{\frac{\alpha}{2}}$ – $\frac{\alpha}{2}$-квантиль распределения $St(n-2)$.

Проверим гипотезу по записанному критерию для данных из условия. Убрав выбросы из данных посчитаем реализацию статистики $T = -0.142$. Для $n = 22$ и $\alpha = 0.05$ получаем критическую область $U_\alpha = (-\infty, -2.086) \cup (2.086, +\infty)$, статистика в нее не попадает, значит гипотеза не отвергается. Или же посчитав значение $p - value = 0.889$ при уровне значимости $\alpha = 0.05$ также получаем, что гипотеза не отклонется.

Для вычисления мощности критерия при альтернативе H_1 в соответствии с определением достаточно проинтегрировать плотность распределения статистики в условиях истинности H_1 по критической области U_α. Однако в данном случае распредление статистики при ненулевом значении коэффициента корреляции получить нетривиально, поэтому прибегнем к оценке мощности с помощью сэмплирования.

Пусть выборки формируются следующим образом. X сэмплируется из многомерного нормального распредления:

$$X \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & r\\ r & 1\end{pmatrix}\right],$$

где в данном случае r – коэффициент корреляции компонент случайного вектора. Первая компонента вектора X относится к первой выборке, вторая – ко второй. Процедура повторяется $n = 22$ раз, где n – мощность выборки. Очевидно, выборки, сэмплируемые таким образом, не являются независимыми, более того существует монотонная зависимость между коэффицентами r и коэффициентом корреляции Спирмана $\rho_{X_1X_2}$ (здесь по сути генерируются нормальные сгущения, для которых этот эффект имеет место).

Перебирая по сетке значения r и сэмплируя указанным способом выборки, вычисляем значение $\rho_{X_1X_2}$, а так же значение статистики и проверяем отклоняется ли гипотеза. Повторяем данную процедуру $N = 10000$ раз и подсчитываем долю неотвергнутых гипотез для каждого узла сетки, так получим оценку для вероятности ошибки второго рода $\hat{\beta} \approx P(H_0|H_1)$, свою для каждой точке исходной сетки. Окончательно мощность критерия оценочна равна $1 - \hat{\beta}$ для каждого узла соотвественно.

Так может быть получена зависимость мощности от r. Оценим истинное значение ρ, как среднее значение $\rho_{X_1X_2}$, рассчитанное по многим сэмплам из указанного многомерного нормального распределения. По данному значению $\hat{\rho}$ вычисляется оценка истинного значения статистики. Необходимо отметить, что $\hat{\rho}$ является плохой оценкой, как минимум, потому что обладает достаточно высокой дисперсией.

Результаты эксперимента представлены на графиках.

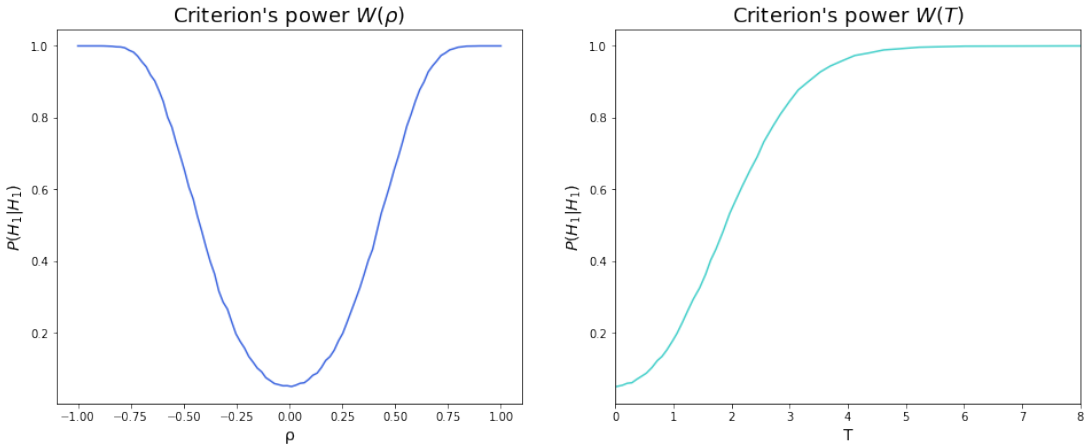

Рис. 1: Зависимость мощности критерия от $\hat{\rho}$ и $T(\hat{\rho})$ соответственно