Распределение Уэлча

Аверьянов, Григорян, Романенко

НИУ "Высшая школа экономики"

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Определение

t-критерий Уэлча — тест, основанный на распределении Стьюдента и предназначенный для проверки статистической гипотезы о равенстве математических ожиданий случайных величин, имеющих необязательно равные известные дисперсии. Является модификацией t-критерия Стьюдента. Назван в честь британского статистика Бернарда Льюиса Уэлча.

Предпосылки

Для применения двухвыборочного **t-критерия Стьюдента** необходимо, чтобы две независимые выборки имели нормальное распределение и истинные дисперсии были равны. В случае **t-критерия Уэлча** дисперсии уже могут быть не равны, но должны быть известны. Предпосылка о нормальном распределении данных сохраняется.

Вычисление статистики

Пусть даны две независимые выборки нормально распределённых случайных величин:

$$X_1, ..., X_{n_x} \sim \mathcal{N}(\mu_x, \sigma_x^2)$$

$$Y_1, ..., Y_{n_y} \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

Проверяем следующую нулевую гипотезу о равенстве математический ожиданий:

$$H_0 : \mu_x = \mu_y$$

Пусть нулевая гипотеза верна. Тогда $E(\overline{X} - \overline{Y}) = 0$ и $Var(\overline{X} - \overline{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$. Пусть $\hat{\sigma}_x^2 = \sum_{i=1}^{n_x} \frac{(X_i - \overline{X})^2}{n_x - 1}$ и $\hat{\sigma}_y^2 = \sum_{i=1}^{n_y} \frac{(Y_i - \overline{Y})^2}{n_y - 1}$ - несмещенные оценки дисперсий $\sigma_x^2 \sigma_y^2$ соответственно.

Рассчитаем следующую статистику:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\widehat{Var}(\bar{X} - \bar{Y})}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} * \frac{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}$$

Распределение первой статистики является стандартным нормальным распределением:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim \mathcal{N}(0, 1)$$

Рассмотрим вторую статистику и для дальнейших вычислений назовем её S :

$$S = \frac{\sigma_x^2/n_x + \sigma_y^2/n_y}{\hat{\sigma}_x^2/n_x + \hat{\sigma}_y^2/n_y}$$

Статистика S напоминает случайную величину с распределением χ^2 , поделенную на степень свободы, но таковой не является. Пусть $Z \sim \chi_d^2$ является случайной величиной с распределением хи-квадрат с d степенями свободы. Мы хотим чтобы S была максимально похожа на $\frac{Z}{d} \sim \frac{\chi_d^2}{d}$, тогда приравняем дисперсии данных случайных величин и выразим d:

$$d = \frac{\left(\sigma_x^2 n_x + \frac{\sigma_y^2}{n_y}\right)^2}{\frac{\sigma_x^4}{n_x^2(n_x - 1)} + \frac{\sigma_y^4}{n_y^2(n_y - 1)}}$$

Итог

В конечном итоге имеем при справедливости нулевой гипотезы:

$$t \overset{approx.}{\sim} t_d$$

где d находится как:

$$d = \left\lfloor \frac{\left(\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)^2}{\frac{\sigma_x^4}{n_x^2(n_x - 1)} + \frac{\sigma_y^4}{n_y^2(n_y - 1)}} \right\rfloor$$

При достаточно больших объемах выборок мы можем воспользоваться нормальной аппроксимацией:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} \xrightarrow[n_x, n_y \to \infty]{} \mathcal{N}(0, 1)$$

Код

Язык	Код
Python	scipy.stats.ttest_ind(a, b, equal_var=False)
R	t.test(data1, data2, alternative="two.sided var.equal=FALSE)
MATLAB	ttest2(data1, data2, 'Vartype', 'unequal')
Excel 2010-	TTEST(array1, array2, tails, type)
Excel 2010+	T.TEST(array1, array2, tails, type)

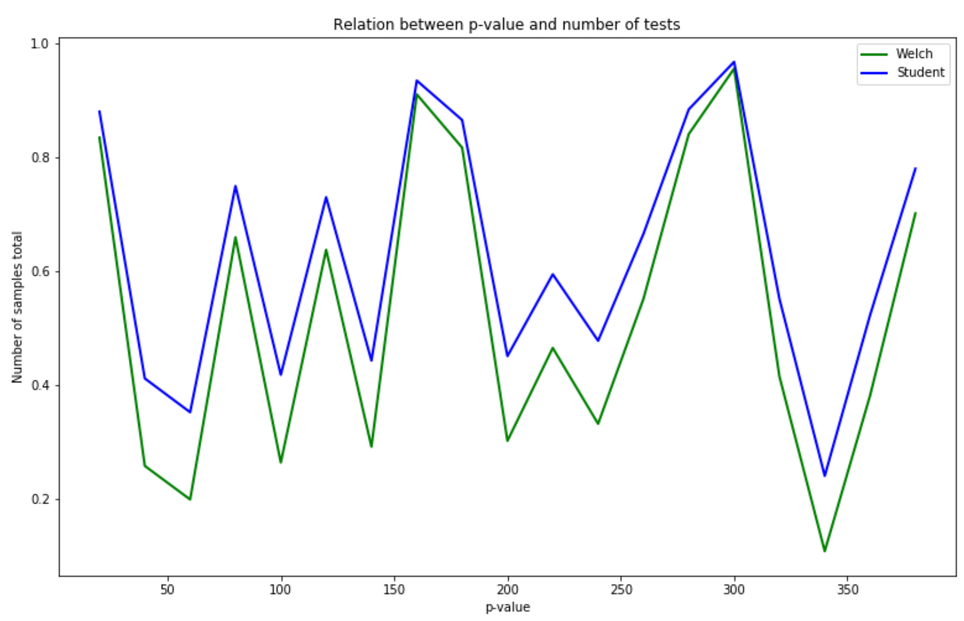

Рис. 1: Relation between p-value and number of tests.

Замечания по использованию

The unequal variance t-test has no performance benefits over the Student's t-test when the underlying population variances are equal. Hence, you might consider that an effective way to conduct your analysis would be to perform an initial test for homogeneity of variance and then perform either a Student's t-test when the variances are equal or an unequal variance t-test when they are not. The problem with this flexible approach is that the combination of this preliminary test plus whichever of the subsequent tests is ultimately used controls Type I error rates less well than simply always performing an unequal variance t-test on every occasion (Gans 1992; Moser and Stevens 1992), this is one reason why it is generally unwise to decide whether to perform one statistical test on the basis of the outcome of another (Zimmerman 2004 and references therein). There are further reasons for not recommending preliminary tests of variances (e.g., Markowski CA and Markowski EP 1990; Quinn and Keough 2002, p. 42).

Ссылки

ну лучше биб техом подгрузить если нужны какие-то