ПРИКЛАДНАЯ СТАТИСТИКА

Равенство средних, неизвестные равные дисперсии

Фишер всё равно бы сумел открыть всё это сам.

Уильям Госсет (Стьюдент) о своих открытиях в статистике

Равенство средних, неизвестные равные дисперсии

У нас есть две независимые выборки из нормально распределённых генеральных совокупностей:

$$X_1, \dots, X_{n_x} \sim iid\ \mathcal{N}(\mu_x, \sigma^2),$$
$$Y_1, \dots, Y_{n_y} \sim iid\ \mathcal{N}(\mu_y, \sigma^2).$$

Дисперсии неизвестны, но одинаковы. Нужно построить доверительный интервал для разности средних на уровне значимости α. Если бы дисперсия была известна, мы бы использовали для этого статистику

$$z = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}}} \sim \mathcal{N}(0, 1).$$

Однако дисперсия неизвестна и её нужно оценить. В связи с этим возникает вопрос, каким будет распределение у новой статистики, где σ^2 заменяется на оценку, посчитанную по нашим двум выборкам.

Если мы по каждой выборке посчитаем оценки дисперсий s_x^2 и s_y^2, то по теореме Фишера

$$\frac{(n_x-1)\cdot s_x^2}{\sigma^2} \sim \chi^2_{n_x-1} \qquad\qquad \frac{(n_y-1)\cdot s_y^2}{\sigma^2} \sim \chi^2_{n_y-1}.$$

Распределение хи-квадрат устойчиво относительно суммирования, то есть если Y_1, Y_2 независимы, и $Y_1 \sim \chi^2_{k_1}$, а $Y_2 \sim \chi^2_{k_2}$, то

$$Y_1 + Y_2 \sim \chi^2_{k_1+k_2}.$$

Получается, что по свойствам хи-квадрат

$$W = \frac{(n_x-1)\cdot s_x^2}{\sigma^2} + \frac{(n_y-1)\cdot s_y^2}{\sigma^2} \sim \chi^2_{n_x+n_y-2}.$$

Попробуем воспользоваться тем же самым приёмом, с помощью которого мы строили t-статистику для проверки гипотезы о среднем. Выписанная для разницы средних z- статистика нормально распределена, но дисперсия в знаменателе неизвестна. Попробуем поделить её на величину W таким образом, чтобы неизвестная дисперсия заменилась на её оценку, а в знаменателе оказалось хи-квадрат распределение. Тогда итоговая статистика будет иметь распределение Стьюдента

$$\frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n_x+n_y-2}}{(n_x+n_y-2)}}} = t(n_x+n_y-2).$$

Получается, что

$$t = \frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma^2}{n_x}+\frac{\sigma^2}{n_y}}} : \sqrt{\frac{[(n_x-1)\cdot s_x^2+(n_y-1)\cdot s_y^2]/\sigma^2}{(n_x+n_y-2)}} \sim t(n_x+n_y-2).$$

Общая выборочная дисперсия будет считаться по формуле

$$s^2 = \frac{1}{n_x+n_y-2}\cdot\left[(n_x-1)\cdot s_x^2+(n_y-1)\cdot s_y^2\right].$$

Две степени свободы расходуются на оценку двух средних. Упростим выражение

$$t = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma^2 \cdot (n_y + n_x)}{n_x \cdot n_y}}} : \sqrt{\frac{s^2}{\sigma^2}} = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s^2 \cdot (n_x + n_y)}{n_x \cdot n_y}}} \sim t(n_x + n_y - 2).$$

Получилось! Когда мы поделили z- статистику на функцию от W, неизвестная дисперсия заменилась на её оценку, а итоговая случайная величина имеет известное нам распределение Стьюдента. Эту случайную величину мы можем использовать для строительства доверительных интервалов и проверки гипотез.