Выборочное среднее, вариация и стандартное отклонение

Если имеем n наблюдений $x_1, x_2, x_3, \ldots x_n$, *из некоторой генеральной совокупности, то* **выборочное среднее** *или* **среднее** *равно*

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n)$$

или $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Вариация определяется как среднеквадратическое отклонение от среднего. **Выборочная вариация** по n наблюдениям $x_1, x_2, x_3, \ldots x_n$ определяется как

$$s^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

или $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

При условии независимости выборки оценки $\bar{x}, s$ являются несмещенными оценками матожидания и дисперсии.

Минимизация среднеквадратического отклонения **Линия, которая минимизирует сумму квадратов отклонений называется** линией наилучшего приближения или линией регресии. Сумма квадратов отклонений от линии наилучшего приближения называется результирующей ошибкой приближения. **.**

Нахождение линии наилучшего приближения

Имеем набор данных, представляющих набор пар (x_i, y_i) и цель -найти линию

$$y = a + bx,$$

ошибка приближения в точке (x_i, y_i) равна $y_i-(a+bx_i)$ и сумма квадратов ошибок

$$D = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2.$$

Для нахождения величин a, b минимизирующих D, обозначаем их $\hat{a}, \hat{b}$, можно воспользоваться равенством нулю производных или замечательным равенством:

$$D = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2 = n(\bar{y} - a - b\bar{x})^2 + (bS_x - \frac{S_{xy}}{S_x})^2 + (S_y^2 - \frac{S_{xy}^2}{S_x^2}) \qquad (*)$$

где

$$S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad S_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Отметим, что S_y^2 называется **“полной вариацией по *y*”**, S_x^2 называется **“полной вариацией по x”** и полные вариации по x и y связаны соотношениями:

$$s_x^2 = S_x^2/(n-1), \quad \text{и} \quad s_y^2 = S_y^2/(n-1).$$

Из этого соотношения находим
a,b**, минимизирующие** D.

$$D = n(\bar{y} - a - b\bar{x})^2 + (bS_x - \frac{S_{xy}}{S_x})^2 + (S_y^2 - \frac{S_{xy}^2}{S_x^2}).$$

Для квадратичного члена минимум равен нулю, для второго слагаемого имеем

$$bS_x - \frac{S_{xy}}{S_x} = 0 \quad \text{или} \quad \boxed{\hat{b} = \frac{S_{xy}}{S_x^2},}$$

Для первого слагаемого

$$\bar{y} - a - \hat{b}\bar{x} = 0 \quad \text{или} \quad \boxed{\hat{a} = \bar{y} - \hat{b}\bar{x}.}$$

Нужна мера насколько хорошо линейное приближение. Эта мера называется **корреляция**.

Основные требования к определению.

i. $r = \pm 1$ для идеального согласования с прямой линией.
Подставив выражения для оптимальной линейной регрессии a,b в выражение для суммы квадратов отклонений получаем:

$$\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2 = S_y^2 - \frac{S_{xy}^2}{S_x^2}.$$

Это выражение равно 0 когда сумма квадратов откло 0.
То есть $r = \pm 1$ эквивалентно условию:

$$S_y^2 - \frac{S_{xy}^2}{S_x^2} = 0 \text{ или } \frac{S_{xy}^2}{S_x^2 S_y^2} = 1.$$

ii. $r=0$ если y не зависит от x. То есть когда $b=0$.

$$\hat{b} = \frac{S_{xy}}{S_x^2}, \textbf{ и следовательно } r=0 \text{ эквивалентно } \hat{b} = \frac{S_{xy}}{S_x^2} = 0.$$

Таким образом очевидный выбор коэффициент корреляции

$$\boxed{r = \frac{S_{xy}}{S_x S_y}.}$$

Заметим, что выборочные вариации для x и y связаны следующим образом:

$s_x^2 = S_x^2/(n-1)$, $s_y^2 = S_y^2/(n-1)$. Далее $s_{xy} = S_{xy}/(n-1)$.
Таким образом имеем альтернативную формулу:

$$\boxed{r = \frac{s_{xy}}{s_x s_y}.}$$

Фактически можно использовать обе формулы, не путая их.

Заметим также, что $S_y^2 - \frac{S_{xy}^2}{S_x^2}$ это сумма квадратичных ошибок, однако если x и y не коррелированы, то наблюдения x не дают никакой информации об y.

На практике часто используются другие типы соотношений для наилучшего согласования переменных независимая переменная x отклик y.

Линейная	$y = a + bx$
Степенная	$y = ax^b$
Экспоненциальная	$y = ae^{bx}$

Нахождение экспоненциального и степенного согласования

Экспоненциальное Если $y = ae^{bx}$ то

$$ln\ y = ln\ a + bx$$

То есть $ln(y)$ есть линейная функция x с константой $ln(a)$ и наклоном b.

Находим линию наилучшего согласования для *ln y* и *x*

b -это наклон линии регресии a = задается как $e^{constant}$

Степенная Если $y = ax^b$ то

$$ln\ y = ln\ a + b\ ln\ x$$

ln(y) есть линейная функция ln(x) с константой ln(a) и наклоном b.

Находим линию наилучшего приближения для *ln y* и *ln(x)*

b это наклон a задается как $e^{constant}$