Поставим следующую задачу: произведено n взаимно независимых экспериментов с двумерной случайной величиной (X, Y). Получены пары чисел (X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n). По имеющимся данным определить ковариацию $\mathrm{cov}(X,Y)=E((X-EX)\cdot(Y-EY))=$ $=E(XY)-(EX)\cdot(EY)$.

Напомним в очередной раз, что это невозможно. И единственное, что удаётся сделать, это найти оценку, то есть формулу, определяющую значение cov(X,Y) по числам X_1, Y_1, X_2, Y_2, … , X_n ,Y_n наилучшим образом с некоторой точки зрения. Самое простое решение заключается в том, что формулы берутся совпадающими с соответствующими им формулами для двумерной дискретной случайной величины, принимающей возможные значения (X_1, Y_1), (X_2, Y_2), … , (X_n, Y_n) с равными вероятностями (равными $1/n$). Таблица распределения такой двумерной дискретной случайной величины имеет вид:

x_i:	X_1	X_2	…	X_n
y_i:	Y_1	Y_2	…	Y_n
p_i:	$1/n$	$1/n$	…	$1/n$

Оценки математических ожиданий, называемые «выборочные средние» известны: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}X_i$ и $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$.

Для ковариации cov(X, Y) хочется предложить оценку $\frac{1}{n}\sum_{i=1}^{n}\left((X_i-EX)(Y_i-EY)\right)$, заменив в ней математические ожидания EX и EY доступными для вычислений средними значениями $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}X_i$ и $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$. Тогда для ковариации cov(X,Y) получится оценка

$$\frac{1}{n}\sum_{i=1}^{n}\left(\left(X_i-\frac{1}{n}\sum_{j=1}^{n}X_j\right)\left(Y_i-\frac{1}{n}\sum_{j=1}^{n}Y_j\right)\right).$$

Проверим эту оценку на выполнение свойства несмещённости: найдём

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left(X_i-\frac{1}{n}\sum_{j=1}^{n}X_j\right)\left(Y_i-\frac{1}{n}\sum_{j=1}^{n}Y_j\right)\right)\right)=\frac{1}{n}\cdot E\left(\sum_{i=1}^{n}\left(\left(X_i-\frac{1}{n}\sum_{j=1}^{n}X_j\right)\left(Y_i-\frac{1}{n}\sum_{j=1}^{n}Y_j\right)\right)\right)=$$

$$=\frac{1}{n}\cdot\sum_{i=1}^{n}E\left(X_iY_i-X_i\frac{1}{n}\left(\sum_{j=1}^{n}Y_j\right)-Y_i\frac{1}{n}\left(\sum_{j=1}^{n}X_j\right)+\frac{1}{n^2}\left(\sum_{j=1}^{n}X_j\right)\left(\sum_{j=1}^{n}Y_j\right)\right)=$$

Это сумма n одинаковых слагаемых, поскольку все X_i имеют одинаковые распределения, такие же, как X_1 и, соответственно, все Y_i имеют одинаковые распределения, такие же, как Y_1.

$$=\frac{1}{n}\cdot nE\left(X_1Y_1-X_1\frac{1}{n}\left(\sum_{j=1}^{n}Y_j\right)-Y_1\frac{1}{n}\left(\sum_{j=1}^{n}X_j\right)+\frac{1}{n^2}\left(\sum_{j=1}^{n}X_j\right)\left(\sum_{j=1}^{n}Y_j\right)\right)=$$

$$=E(X_1Y_1)-E\left(X_1\frac{1}{n}\left(\sum_{j=1}^{n}Y_j\right)\right)-E\left(Y_1\frac{1}{n}\left(\sum_{j=1}^{n}X_j\right)\right)+E\left(\frac{1}{n^2}\left(\sum_{j=1}^{n}X_j\right)\left(\sum_{j=1}^{n}Y_j\right)\right)=$$

$$=E(X_1Y_1)-\frac{1}{n}\cdot E\left(X_1\left(\sum_{j=1}^{n}Y_j\right)\right)-\frac{1}{n}\cdot E\left(Y_1\left(\sum_{j=1}^{n}X_j\right)\right)+\frac{1}{n^2}\cdot E\left(\left(\sum_{j=1}^{n}X_j\right)\left(\sum_{j=1}^{n}Y_j\right)\right)=$$

Теперь нужно понять следующее. Результаты разных испытаний независимые. А математическое ожидание независимых случайных величин равно произведению их математических ожиданий. Например, при j>1 верно $E(X_1Y_j)=EX_1\cdot EY_j$. Поэтому для

продолжения выкладок нужно отделить слагаемые, где есть произведения результатов испытаний с разными индексами от произведений результатов с одинаковыми индексами.

$$= E(X_1Y_1) - \frac{1}{n}\cdot E\left(X_1\left(Y_1 + \sum_{j=2}^{n} Y_j\right)\right) - \frac{1}{n}\cdot E\left(Y_1\left(X_1 + \sum_{j=2}^{n} X_j\right)\right) + \frac{1}{n^2}\cdot E\left(\sum_{j=1}^{n}\sum_{k=1}^{n} X_jY_k\right) =$$

Суммы умножились обычным образом (умножается каждое слагаемое первой суммы на каждое слагаемое второй суммы).

$$= E(X_1Y_1) - \frac{1}{n}\cdot E\left(X_1Y_1 + \sum_{j=2}^{n} X_1Y_j\right) - \frac{1}{n}\cdot E\left(X_1Y_1 + \sum_{j=2}^{n} X_jY_1\right) + \frac{1}{n^2}\cdot E\left(\sum_{j=1}^{n} X_jY_j + \sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n} X_jY_k\right) =$$

В двойной сумме необходимо было отделить друг от друга слагаемые с одинаковыми индексами от слагаемых с разными индексами.

$$= E(X_1Y_1) - \frac{1}{n}\cdot\left(E(X_1Y_1) + E\sum_{j=2}^{n} X_1Y_j\right) - \frac{1}{n}\cdot\left(E(X_1Y_1) + E\sum_{j=2}^{n} X_jY_1\right) + \frac{1}{n^2}\cdot\left(E\sum_{j=1}^{n} X_jY_j + E\sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n} X_jY_k\right) =$$

$$= E(X_1Y_1) - \frac{1}{n}\cdot\left(E(X_1Y_1) + \sum_{j=2}^{n} E(X_1Y_j)\right) - \frac{1}{n}\cdot\left(E(X_1Y_1) + \sum_{j=2}^{n} E(X_jY_1)\right) + \frac{1}{n^2}\cdot\left(\sum_{j=1}^{n} E(X_jY)_j + \sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n} E(X_jY_k)\right) =$$

$$= E(X_1Y_1) - \frac{1}{n}\cdot\left(E(X_1Y_1) + \sum_{j=2}^{n} E(X_1)E(Y_j)\right) - \frac{1}{n}\cdot\left(E(X_1Y_1) + \sum_{j=2}^{n} E(X_j)E(Y_1)\right) +$$

$$+ \frac{1}{n^2}\cdot\left(\sum_{j=1}^{n} E(X_jY_j) + \sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n} E(X_j)E(Y_k)\right) =$$

В очередной раз обратим внимание, что в каждой из сумм все слагаемые одинаковые. В частности, $E(X_j)=E(X_1)$, и $E(Y_k)=E(Y_1)$.

Аккуратно подсчитаем количества слагаемых в каждой из сумм.

$$= E(X_1Y_1) - \frac{1}{n}\cdot\left(E(X_1Y_1) + (n-1)E(X_1)E(Y_1)\right) - \frac{1}{n}\cdot\left(E(X_1Y_1) + (n-1)E(X_1)E(Y_1)\right) +$$

$$+ \frac{1}{n^2}\cdot\left(nE(X_1Y_1) + \left(n^2-n\right)E(X_1)E(Y_1)\right) =$$

$$= E(X_1Y_1) - \frac{2}{n}\cdot E(X_1Y_1) - \frac{2(n-1)}{n}E(X_1)E(Y_1) + \frac{1}{n}\cdot E(X_1Y_1) + \frac{n-1}{n}E(X_1)E(Y_1) =$$

$$= E(X_1Y_1) - \frac{1}{n}\cdot E(X_1Y_1) - \frac{n-1}{n}E(X_1)E(Y_1) = \frac{n-1}{n}\cdot E(X_1Y_1) - \frac{n-1}{n}E(X_1)E(Y_1) =$$

$$= \frac{n-1}{n}\cdot\left(E(X_1Y_1) - E(X_1)E(Y_1)\right) = \frac{n-1}{n}\cdot\operatorname{cov}(X_1,Y_1) = \frac{n-1}{n}\cdot\operatorname{cov}(X,Y) \neq \operatorname{cov}(X,Y)$$

Оценка не прошла тест на несмещённость. В то же время ясно, как исправить эту оценку, чтобы её математическое ожидание было равно cov(X,Y). Нужно умножить прежнюю оценку на дробь $\frac{n}{n-1}$. Таким образом «выборочную ковариацию» разумнее вычислять по формуле

$$\operatorname{cov}(X,Y) = \frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\left(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j\right)\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\left(Y_i - \frac{1}{n}\sum_{j=1}^{n} Y_j\right)\right).$$

Такая оценка уже будет несмещённой.

Обычно расчётчики найдя ковариацию не останавливаются и вычисляют корреляцию

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{DX}\cdot\sqrt{DY}}.$$

Подставив найденные оценки для ковариации и дисперсий получим расчетную формулу для оценки корреляции

$$r_{XY} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)\left(Y_i - \frac{1}{n}\sum_{j=1}^{n}Y_j\right)\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right)} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\left(Y_i - \frac{1}{n}\sum_{j=1}^{n}Y_j\right)^2\right)}}$$

Дробь можно сократить

$$r_{XY} = \frac{\sum_{i=1}^{n}\left(\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)\left(Y_i - \frac{1}{n}\sum_{j=1}^{n}Y_j\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right) \cdot \sum_{i=1}^{n}\left(\left(Y_i - \frac{1}{n}\sum_{j=1}^{n}Y_j\right)^2\right)}}$$