# ОСНОВЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА

## 1. Исходные понятия

При решении прикладных задач в различных областях человеческой деятельности, в том числе и в инженерной практике, исследователь нередко сталкивается с необходимостью установления факта существования функциональных или иных зависимостей между переменными величинами, которые могут быть и случайными. Для подтверждения сказанного рассмотрим несколько простейших примеров.

**Пример 6.1.** Пусть $Y$ — величина износа (в мм) протектора шины на автомобилях определенного типа после 10000 км пробега, $X_1$ — величина нагрузки (в кг) на колесо автомобиля, $X_2$ — тип протектора (используются три типа протекторов). Если установить степень влияния $X_1$ и $X_2$ на $Y$, то можно дать рекомендации по продлению долговечности шины.

**Пример 6.2.** Пусть $Y_1$ — производительность химической установки (в т/ч), $Y_2$ — процент брака готовой продукции. Технолог предполагает, что на переменные $Y_1$ и $Y_2$ влияют в наибольшей степени такие технологические параметры, как: $X_1$ — влажность сырья (в %), $X_2$ — температура в реакторе установки, $X_3$ — содержание примеси (в %).

Как установить степень влияния контролируемых переменных $X_1$, $X_2$, $X_3$ на переменные $Y_1$ и $Y_2$? Если найти вид зависимости $Y_1$ и $Y_2$ от $X_1$, $X_2$, $X_3$, то можно выбрать оптимальный (т.е. наилучший в определенном смысле) технологический режим (при котором, например, процент брака будет минимальным при заданном уровне производительности).

**Пример 6.3.** Пусть $Y$ — успеваемость студентов по некоторой дисциплине (измеряемая, например, средним баллом на экзамене). Деканат проводит обследование студентов данного вуза с целью установления наиболее значимых факторов, влияющих на $Y$. В результате предварительного анализа сделано предположение о том, что этими факторами могут быть: $X_1$ — время, затрачиваемое студентом на самостоятельную работу, $X_2$ — количество пропущенных занятий, $X_3$ — величина стипендии. Существует ли взаимосвязь между факторами $X_1$, $X_2$, $X_3$? В какой степени они оказывают влияние на успеваемость? #

Приведенные примеры далеко не полностью отражают возможные постановки задач рассматриваемого типа. Но даже их поверхностный анализ позволяет отметить следующее.

1. Зависимое переменное $Y$ может быть случайной величиной, даже если переменные $X_1$, ..., $X_p$ таковыми не являются, так как значение $Y$ определяется не только значениями переменных $X_1$, ..., $X_p$, которые исследователь выделил (по его мнению, они являются определяющими), но и многими другими неучтенными факторами, а также ошибками измерений. Это означает, что ***связь*** между $X_1, \ldots, X_p$ и $Y$ является не функциональной, а ***стохастической*** — изменение переменных $X_1, \ldots,$ $X_p$ влияет на значения переменного $Y$ через изменение закона распределения случайной величины $Y$.

2. Некоторые переменные могут иметь количественный характер, а некоторые — качественный (см. пример 6.1).

3. Нас может интересовать либо зависимость переменного $Y$ от переменных $X_1$, ..., $X_p$, либо взаимозависимость между несколькими переменными (не обязательно между всеми). Так, в примере 6.3 может существовать взаимозависимость между переменными $X_1$, $X_2$ и $X_3$.

Перечисленные особенности приводят к различным постановкам задач статистического исследования зависимостей, которые упрощенно можно классифицировать следующим образом:

1) ***задачи корреляционного анализа*** — задачи исследования наличия взаимосвязей между отдельными группами переменных;

2) ***задачи регрессионного анализа*** — задачи, связанные с установлением аналитических зависимостей между переменным $Y$ и одним или несколькими переменными $X_1, \ldots, X_p$, которые носят количественный характер;

3) ***задачи дисперсионного анализа*** — задачи, в которых переменные $X_1, \ldots, X_p$ имеют качественный характер, а исследуется и устанавливается степень их влияния на переменное $Y$.

Кроме перечисленных типов задач выделяют и многие другие. Так, ковариационный анализ рассматривает одновременно и количественные и качественные переменные $X_1$, $\ldots, X_p$, ***конфлюентный анализ**** обобщает регрессионный на тот случай, когда переменные $X_1, \ldots, X_p$ и $Y$ измеряют с ошибками, ***факторный анализ***** служит для выделения из множества исследуемых переменных $X_1, \ldots, X_p$ наиболее значимых***.

Для удобства дальнейших рассуждений обратимся к так называемой модели „черного ящика" (рис. 6.1) как наиболее общей модели любой реальной системы, ассоциированной с понятием отображения $f: \vec{X} \to \vec{Y}$. На вход „черного ящика" поступает входной сигнал — вектор $\vec{X}$, который посредством отображения $f$ преобразуется в выходной сигнал — вектор $\vec{Y}$. При этом, в соответствии со сложившийся терминологией,

$\vec{X} = (X_1, \ldots, X_p)$ — вектор ***входных переменных***, или вектор

***факторов***; $\vec{Y} = (Y_1, \ldots, Y_m)$ — вектор ***выходных переменных***, или вектор ***откликов***; $\varepsilon = \vec{Y} - f(\vec{X})$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ — вектор ***случайных ошибок***, т.е. случайных переменных, отражающих влияние на переменные $Y_i$, $i = \overline{1, m}$, неучтенных факторов, а также случайных ошибок измерений анализируемых показателей.

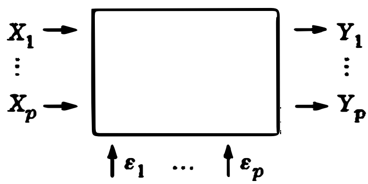


**Рис. 6.1**

При проведении корреляционного анализа исследователь должен уметь:

а) выбрать показатель стохастической связи анализируемых переменных;

б) оценить его значение по имеющимся *экспериментальным данным*, т.е. найти его *точечную* и *интервальную оценки*;

в) проверить *статистическую гипотезу* о том, что значение показателя стохастической связи значимо отличается от нуля.

Ниже дано описание методов и моделей, используемых для решения перечисленных задач.

## 2. Анализ парных связей

**Выбор показателя связи.** Для начала рассмотрим задачу выбора показателя *стохастической связи* между двумя случайными величинами* $\xi$ и $\eta$, реализации которых будем обозначать соответственно через $x$ и $y$.

**Пример 6.4.** Пусть случайный вектор $(\xi,\eta)$ имеет нормальный закон распределения с математическим ожиданием $\mu=(\mu_1,\mu_2)$ и ковариационной матрицей

$$\Sigma=\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

где $\sigma_1^2$ и $\sigma_2^2$ — дисперсии случайных величин $\xi$ и $\eta$ соответственно, а $\rho$ — коэффициент корреляции между $\xi$ и $\eta$.

В этом случае условная плотность распределения случайной величины $\eta$ при условии, что $\xi=x$,

$$p(y\,|\,x)=\frac{1}{\sigma_{\eta/x}\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_{\eta/x}}{\sigma_{\eta/x}}\right)^2\right)$$

является плотностью нормального распределения с параметрами $\mu_{\eta/x}$ (условное математическое ожидание) и $\sigma^2_{\eta/x}$ (условная дисперсия $\eta$) при значении $\xi=x$, которые связаны с параметрами исходного двумерного распределения следующим образом:

$$\mathbf{M}(\eta\,|\,\xi=x)=\mu_{\eta/x}=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \tag{6.1}$$

$$\mathbf{D}(\eta\,|\,\xi=x)=\sigma^2_{\eta/x}=\sigma_2^2(1-\rho^2). \tag{6.2}$$

В рассматриваемом случае линия регрессии является прямой, а условная дисперсия не зависит от $x$. #

Если закон распределения случайного вектора $(\xi,\eta)$ не является нормальным, то характер изменения условного математического ожидания $\mathbf{M}(\eta\,|\,\xi=x)=f(x)$ может быть и нелинейным, причем, чем меньше условная дисперсия $\mathbf{D}(\eta\,|\,\xi=x)$, тем меньше при различных значениях $x$ рассеяны возможные значения случайной величины $\eta$ относительно линии регрессии $\mathbf{M}(\eta\,|\,\xi=x)=f(x)$ (рис. 6.2). Функцию $f(x)=\mathbf{M}(\eta\,|\,\xi=x)$ называют функцией регрессии, или регрессией.

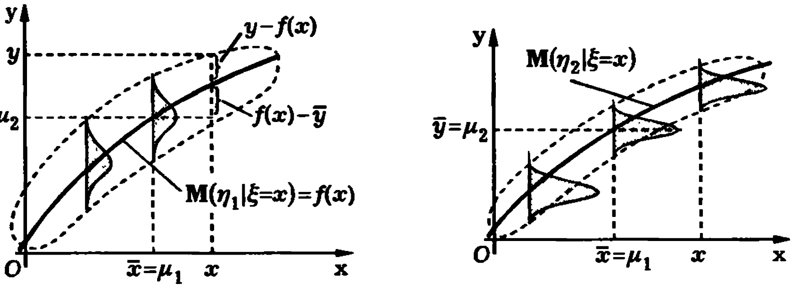


Рис. 6.2

Обозначим $\mathbf{M}\eta = \mu$, $\mathbf{D}\eta = \sigma_\eta^2$. Отклонение $y-\mu$ возможных значений $\eta$ от $\mu$ складывается из двух слагаемых (см. рис. 6.2):

$$y-\mu = (f(x)-\mu) + (y - f(x)), \tag{6.3}$$

где $f(x)-\mu$ — отклонение функции регрессии $f(x)$ в точке $x$ от математического ожидания $\mu$; $y-f(x)$ — отклонение возможного значения $\eta$ от значения функции регрессии в точке $x$.

Покажем, что рассеяние $\sigma_\eta^2$ случайной величины $\eta$ относительно ее математического ожидания есть сумма двух слагаемых, а именно: математического ожидания квадрата отклонения $\eta$ от ее условного математического ожидания $f(\xi)$ и математического ожидания квадрата отклонения $f(\xi)$ от $\mu$.

Действительно ,

$$\mathbf{M}(f(\xi)) = \mathbf{M}(\mathbf{M}(\eta|\xi)) = \mathbf{M}\eta = \mu,$$

$$\mathbf{D}\eta = \sigma_\eta^2 = \mathbf{M}(\eta-\mu)^2 = \mathbf{M}\Big((\eta - f(\xi)) + (f(\xi)-\mu)\Big)^2 =$$

$$= \mathbf{M}(\eta - f(\xi))^2 + 2\mathbf{M}((\eta - f(\xi))(f(\xi)-\mu)) + \mathbf{M}(f(\xi)-\mu)^2 =$$

$$= \mathbf{M}(\eta - f(\xi))^2 + \mathbf{M}(f(\xi)-\mu)^2,$$

так как $\mathbf{M}\big((\eta - f(\xi))(f(\xi)-\mu)\big) = 0.$

Докажем последнее равенство для непрерывных случайных величин $\xi$ и $\eta$, предполагая, что их совместная плотность распределения $p(x,y)$ в $\mathbb{R}^2$ не обращается в нуль:

$$\mathbf{M}\big((\eta - f(\xi))(f(\xi)-\mu)\big) = \iint\limits_{-\infty}^{\infty} (y - f(x))(f(x)-\mu)\,p(x,y)\,dx\,dy =$$

$$= \int\limits_{-\infty}^{\infty} (f(x)-\mu)\,p_\xi(x)\,dx \int\limits_{-\infty}^{\infty} (y - f(x))\frac{p(x,y)}{p_\xi(x)}\,dy =$$

$$= \int\limits_{-\infty}^{\infty} (f(x)-m)\,p_\xi(x)\,dx \left( \int\limits_{-\infty}^{\infty} y\frac{p(x,y)}{p_\xi(x)}\,dy - f(x)\int\limits_{-\infty}^{\infty} \frac{p(x,y)}{p_\xi(x)}\,dy \right) = 0,$$

так как

$$\int\limits_{-\infty}^{\infty} y\frac{p(x,y)}{p_\xi(x)}\,dy = f(x), \qquad \int\limits_{-\infty}^{\infty} \frac{p(x,y)}{p_\xi(y)}\,dy = 1.$$

Таким образом, если воспользоваться обозначениями

$$\sigma_f^2 = \mathbf{D}\,f(\xi) = \mathbf{M}\big(f(\xi)-\mu\big)^2, \qquad \overline{\sigma}_\eta^2 = \mathbf{M}\big(\eta - f(\xi)\big)^2,$$

то полученный результат может быть представлен в виде

$$\sigma_\eta^2 = \sigma_f^2 + \overline{\sigma}_\eta^2. \tag{6.4}$$

Из равенства (6.4) следует, что связь между $\xi$ и $\eta$ тем теснее, чем меньше слагаемое $\overline{\sigma}_\eta^2$ или чем больший вклад в дисперсию $\sigma_\eta^2$ вносит слагаемое $\sigma_f^2$, порожденное функцией регрессии $f(x) = \mathbf{M}(\eta)\xi = x$. Тем самым мы приходим к понятию общей характеристики степени тесноты связи — ***корреляционному отношению*** переменного $\eta$ по переменному $\xi$:

$$r_{\eta\xi} = \sqrt{\frac{\sigma_f^2}{\sigma_\eta^2}} = \sqrt{1 - \frac{\overline{\sigma}_\eta^2}{\sigma_\eta^2}}. \tag{6.5}$$

Непосредственно из (6.5) следует, что всегда выполняется неравенство

$$0 \leqslant r_{\eta\xi}^2 \leqslant 1, \tag{6.6}$$

причем равенство $r_{\eta\xi}^2 = 0$ означает, что с изменением $\xi$ вариация функции регрессии $f(x)$ полностью отсутствует. Другими словами, случайные величины $\xi$ и $\eta$ являются независимыми. В этом случае линия регрессии есть горизонтальная прямая. Равенство $r_{\xi\eta}^2 = 1$ будет иметь место, если $\overline{\sigma}_\eta^2 = \mathbf{M}\big(\eta - f(\xi)\big)^2 = 0$, т.е. если $\eta$ и $\xi$ связаны функциональной зависимостью $\eta = f(\xi)$.

Аналогично определяется корреляционное отношение $r_{\xi\eta}$ переменной $\xi$ по $\eta$.

**Замечание 6.1.** Между $r_{\eta\xi}$ и $r_{\xi\eta}$ нет какой-либо простой зависимости. Возможны ситуации, в которых один из этих показателей принимает нулевое значение, в то время как другой равен единице. Пусть, например, $\eta = \xi^2$, а $\xi$ принимает следующие значения: $-1$, $0$, $1$ с вероятностями $1/3$ каждое. В этом случае $r_{\eta\xi} = 1$, $r_{\xi\eta} = 0$ (в силу симметричности параболы относительно оси значений $\eta$ и симметричности распределения $\xi$). #

Итак, решение задачи выбора показателя стохастической связи между двумя случайными величинами $\xi$ и $\eta$ для самой общей ситуации, когда закон распределения вектора $(\xi, \eta)$ является произвольным, найдено — таким показателем являются корреляционные отношения $r_{\eta\xi}$ и $r_{\xi\eta}$.

Выясним, какую роль играет такой показатель связи между случайными величинами $\xi$ и $\eta$, как коэффициент корреляции $\rho$:

$$\rho = \frac{\mathbf{M}\big((\xi - \mathbf{M}\xi)(\eta - \mathbf{M}\eta)\big)}{\sigma_1\sigma_2}, \tag{6.7}$$

где $\sigma_1 = \sqrt{\mathbf{D}\xi}$, $\sigma_2 = \sqrt{\mathbf{D}\eta}$, $\mathbf{M}\big((\xi - \mathbf{M}\xi)(\eta - \mathbf{M}\eta)\big)$ — второй смешанный момент случайного вектора $(\xi, \eta)$.

Напомним, что случайные величины $\xi$ и $\eta$ называют некоррелированными, если $\rho = 0$, и коррелированными при $\rho \neq 0$.

Известно, что из независимости случайных величин $\xi$ и $\eta$ следует их некоррелированность, однако обратное утверждение в общем случае неверно.

Если случайный вектор $(\xi,\eta)$ имеет нормальный закон распределения, то линия регрессии $\eta$ по $\xi$ (и $\xi$ по $\eta$) является прямой (см. пример 6.4), т.е. коэффициент корреляции $\rho$ может служить мерой связи между $\xi$ и $\eta$. Для нормального закона распределения на основании (6.2) и (6.5) имеем

$$r^2_{\eta\xi} = r^2_{\xi\eta} = \rho^2.$$

Действительно, из (6.2) получаем, что условная дисперсия $\eta$ не зависит от значений случайной величины $\xi$, и, следовательно,

$$\sigma^2_\eta(1-\rho^2) = \mathbf{D}(\eta\,|\,\xi) = \mathbf{M}\Big(\big(\eta - \mathbf{M}(\eta\,|\,\xi)\big)^2\,|\,\xi\Big) =$$

$$= \mathbf{M}\Big(\big(\eta - f(\xi)\big)^2\,|\,\xi\Big) = \mathbf{M}\big(\eta - f(\xi)\big)^2 = \bar{\sigma}^2_\eta.$$

Наконец, учитывая (6.5) и полученный результат, приходим к равенству $r^2_{\eta\xi} = \rho^2$. Аналогично можно доказать равенство $r^2_{\xi\eta} = \rho^2$. Таким образом, корреляционные отношения совпадают между собой и с абсолютной величиной коэффициента корреляции $\rho$. При этом равенство $|\rho| = 1$ означает линейную функциональную зависимость между $\xi$ и $\eta$, а равенство $\rho = 0$ свидетельствует об их линейной независимости.

Понятно, что рассмотренными свойствами двумерного нормального закона не могут обладать все двумерные законы распределения или хотя бы их большая часть. Поэтому в общем случае не имеет смысла использование коэффициента корреляции $\rho$ как меры взаимосвязи случайных величин $\xi$ и $\eta$.

В общем случае показатели $r^2_{\eta\xi}$ и $\rho^2$ связаны неравенствами

$$0 \leqslant \rho^2 \leqslant r^2_{\eta\xi} \leqslant 1. \tag{6.8}$$

При этом возможны следующие варианты:

а) $\rho^2 = 0$, если $\xi$ и $\eta$ независимы, но обратное (в общем случае) неверно;

б) $\rho^2 = r^2_{\eta\xi} = 1$ тогда и только тогда, когда имеется строгая линейная функциональная зависимость $\eta$ от $\xi$;

в) $\rho^2 \leqslant r^2_{\eta\xi} = 1$ тогда и только тогда, когда имеется строгая нелинейная функциональная зависимость $\eta$ от $\xi$;

г) $\rho^2 = r^2_{\eta\xi} < 1$ тогда и только тогда, когда регрессия $\eta$ по $\xi$ строго линейна, но нет функциональной зависимости;

д) $\rho^2 < r^2_{\eta\xi} < 1$ указывает на то, что не существует функциональной зависимости, а некоторая нелинейная кривая регрессии „подходит" лучше, чем „наилучшая" прямая линия.

Итак, в качестве показателя стохастической связи между двумя случайными количественными переменными $\xi$ и $\eta$ следует выбрать корреляционное отношение $r_{\eta\xi}$ (или $r_{\xi\eta}$), если закон распределения вектора $(\xi, \eta)$ вызывает сомнение. Если же можно с большой степенью уверенности считать закон распределения вектора $(\xi, \eta)$ нормальным, то вместо корреляционного отношения следует использовать коэффициент корреляции $\rho$.

**Оценка показателя связи по выборочным данным.** После выбора показателя стохастической связи задача корреляционного анализа, как уже отмечалось в **6.1**, состоит в нахождении его *оценки* (*точечной* и *интервальной*), а также в проверке *статистической гипотезы* о значимом отличии его от нуля на основе *экспериментальных данных*.

Пусть в результате эксперимента получены *n выборочных значений* случайного вектора $(\xi, \eta)$, которые будем записывать в виде

$$(x_i, y_i), \quad i = \overline{1, n}. \tag{6.9}$$

При изучении корреляционной зависимости двух случайных величин $(\xi, \eta)$ по *выборке* $(x_i, y_i)$, $i = \overline{1, n}$, общую картину их взаимной изменчивости можно получить, изобразив на координатной плоскости все точки. Это изображение называют ***корреляционным полем***.

Уже по виду корреляционного поля можно иногда сделать вывод о наличии и характере связи между случайными величинами $\xi$ и $\eta$. Так, на рис. 6.3, *а* выборочные точки $(x_i, y_i)$ лежат внутри некоторого эллипса (эллипса рассеяния) с осями, параллельными координатным. Следовательно, с изменением, например, $\xi$ величина $\eta$ не будет менять своего условного распределения, т.е. $\xi$ и $\eta$, по-видимому, некоррелированы. Напротив, на рис. 6.3, *б* видно, что условное математическое ожидание $\mathrm{M}(\eta|\xi = x) = f(x)$ имеет линейный характер изменения, и, значит, следует ожидать, что коэффициент корреляции $\rho$ близок к единице. На рис. 6.3, *в* расположение точек $(x_i, y_i)$ говорит о наличии нелинейного характера изменения $f(x)$, и, следовательно, коэффициент корреляции может оказаться близким к нулю, а корреляционное отношение $r_{\eta\xi}$ — близким к единице.

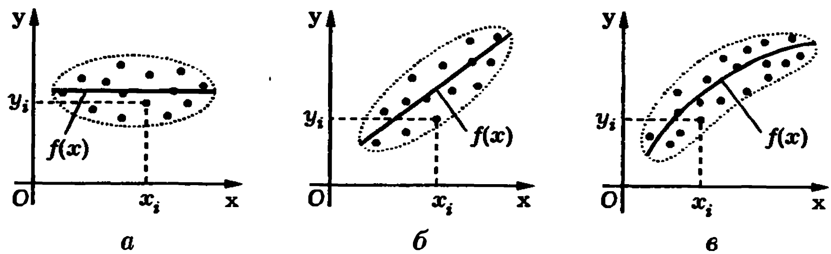


**Рис. 6.3**

Следует отметить, что в том случае, когда среди $x_i$ есть повторяющиеся с частотой $n_i$ значения, выборочные значения представляют в виде

$$(x_i,\ y_{ij}), \quad j = \overline{1, n_i},\ i = \overline{1, m}, \quad \sum_{i=1}^{m} n_i = n. \tag{6.10}$$

Если выборочные значения сгруппированы по каждой из переменных, т.е. значения $x_i$ разделены на $m$ групп, а значения $y_i$ — на $l$ групп, то выборочные значения представляют в виде

$$(x_i,\ y_j,\ n_{ij}), \quad i = \overline{1, m},\ j = \overline{1, l}, \quad \sum_{i,j} n_{ij} = n, \tag{6.11}$$

или в виде ***корреляционной таблицы***, в каждой клетке которой указывают число $n_{ij}$ попавших в нее выборочных значений, причем сумма всех этих значений равна $n$ (табл. 6.1).

*Таблица 6.1*

| Значения $\xi$ | Значения $\eta$ | | | | |
|---|---|---|---|---|---|
| | $y_1$ | ... | $y_j$ | ... | $y_l$ |
| $x_1$ | $n_{11}$ | ... | $n_{1j}$ | ... | $n_{1l}$ |
| ... | ... | ... | ... | ... | ... |
| $x_i$ | $n_{i1}$ | ... | $n_{ij}$ | ... | $n_{il}$ |
| ... | ... | ... | ... | ... | ... |
| $x_m$ | $n_{m1}$ | ... | $n_{mj}$ | ... | $n_{ml}$ |

## 6.3. Анализ коэффициента корреляции

**Точечная оценка показателя $\rho$.** Пусть *экспериментальные данные* представлены в форме (6.9). Тогда $\widehat{\rho}$ — значение *точечной оценки* коэффициента корреляции $\rho$ — вычисляют по формуле

$$\widehat{\rho}=\frac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}}. \qquad (6.12)$$

**Пример 6.5.** Вычислим значение $\widehat{\rho}$ для пары случайных величин $(\xi,\ \eta)$, где $\xi$ — рост (в см), а $\eta$ — масса тела (в кг) наугад выбранного студента-первокурсника. Выборка объема $n=15$ представлена в табл. 6.2.

Чтобы оценить показатель $\rho$ связи двух случайных величин, сначала найдем выборочные средние этих величин:

$$\overline{x}=\frac{1}{15}\sum_{i=1}^{15}x_i=\frac{2620}{15}=173{,}3; \qquad \overline{y}=\frac{1}{15}\sum_{i=1}^{15}y_i=\frac{945}{15}=63{,}1.$$

*Таблица 6.2*

| Номер наблюдения | Рост, см $x_i$ | $x_i - \overline{x}$ | Масса тела, кг $y_i$ | $y_i - \overline{y}$ |
|---|---|---|---|---|
| 1 | 165 | −8,3 | 72,9 | 9,8 |
| 2 | 171 | −2,3 | 48,4 | −14,7 |
| 3 | 182 | 8,7 | 66,3 | 3,2 |
| 4 | 165 | −8,3 | 64,1 | 1,0 |
| 5 | 183 | 9,7 | 62,7 | −0,4 |
| 6 | 180 | 6,7 | 76,0 | 12,9 |
| 7 | 183 | 9,7 | 73,8 | 10,7 |
| 8 | 166 | −7,3 | 50,6 | −12,5 |
| 9 | 173 | −0,3 | 52,3 | −10,8 |
| 10 | 172 | −1,3 | 56,5 | −6,6 |
| 11 | 174 | 0,7 | 66,8 | 3,7 |
| 12 | 170 | −3,3 | 61,6 | −1,5 |
| 13 | 164 | −9,3 | 72,8 | 9,7 |
| 14 | 168 | −5,3 | 52,6 | −10,5 |
| 15 | 184 | 10,7 | 68,6 | 5,5 |
| Σ | 2600 | | 945 | |

Затем определяем суммы

$$\sum_{i=1}^{15}(x_i - \overline{x})^2 = 747{,}33; \qquad \sum_{i=1}^{15}(y_i - \overline{y})^2 = 1171{,}4;$$

$$\sum_{i=1}^{15}(x_i - \overline{x})(y_i - \overline{y}) = 293{,}3.$$

Таким образом, $\widehat{\rho} = \dfrac{293{,}3}{\sqrt{747{,}33 \cdot 1171{,}4}} = 0{,}313.$

**Замечание 6.2.** Если экспериментальные данные представлены в виде (6.10) или (6.11), т.е. сгруппированы по одному или по обоим переменным, то расчетная формула (6.12) для $\widehat{\rho}$ изменяется соответствующим образом. Например, если *выборка* представлена в виде (6.10), то значения оценок $\widehat{\mu}_{11}$, $\widehat{\sigma}_1$ и $\widehat{\sigma}_2$

вычисляют по формулам

$$\widehat{\mu}_{1,1} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x})(\overline{y}_i - \overline{\overline{y}}), \qquad \overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \qquad \overline{\overline{y}} = \frac{1}{m}\sum_{i=1}^{m}\overline{y}_i,$$

$$\widehat{\sigma}_1 = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - \overline{x})^2}, \qquad \widehat{\sigma}_2 = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\overline{y}_i - \overline{\overline{y}})^2}.$$