4. АНАЛИЗ РЕЗУЛЬТАТОВ ПАССИВНОГО ЭКСПЕРИМЕНТА. ЭМПИРИЧЕСКИЕ ЗАВИСИМОСТИ

4.1. Характеристика видов связей между рядами наблюдений

На практике сама необходимость измерений большинства величин вызывается тем, что они не остаются постоянными, а изменяются в функции от изменения других величин. В этом случае целью проведения эксперимента является установление вида функциональной зависимости $\hat{y}$=f(**X**). Для этого должны одновременно определяться как значения **X**, так и соответствующие им значения $\hat{y}$, а задачей эксперимента является установление математической модели исследуемой зависимости. Фактически речь идет об установлении связи между двумя рядами наблюдений (измерений).

Определение связи включает в себя указание вида модели и определение ее параметров. В теории экспериментов независимые параметры **X**=(x_1, ..., x_k) принято называть факторами, а зависимые переменные y – откликами. Координатное пространство с координатами x_1, x_2, ..., x_i, ..., x_k называется факторным пространством. Эксперимент по определению вида функции

$$\hat{y} = f(x), \qquad (4.1)$$

где x – скаляр, называется однофакторным. Эксперимент по определению функции вида

$$\hat{y} = f(\mathbf{X}), \qquad (4.1а)$$

где **X**=(x_1, x_2, ..., x_i, ..., x_k) – вектор – многофакторным.

Геометрическим представлением функции отклика в факторном пространстве является поверхность отклика. При однофакторном эксперименте (k=1) поверхность отклика представляет собой линию на плоскости, при двухфакторном (k=2) – поверхность в трехмерном пространстве.

Связи в общем случае являются достаточно многообразными и сложными. Обычно выделяют следующие виды связей.

Функциональные связи (или зависимости) – это такие связи, когда при изменении величины **X** другая величина y изменяется так, что каждому значению x_i соответствует совершенно определенное (однозначное) значение y_i (рис.4.1,а). Таким образом, если выбрать все условия эксперимента абсолютно одинаковыми, то, повторяя испытания, получим одну и ту же зависимость, т.е. кривые идеально совпадут для всех испытаний.

К сожалению, такие условия в реальности не встречаются. На практике не удается поддерживать постоянство условий (например, физико-химические свойства шихты при моделировании процессов тепломассопереноса в металлургических печах). При этом влияние каждого случайного фактора в отдельности может быть мало, однако в совокупности они существенно могут повлиять на результаты эксперимента. В этом случае говорят о стохастической (вероятностной) связи между переменными.

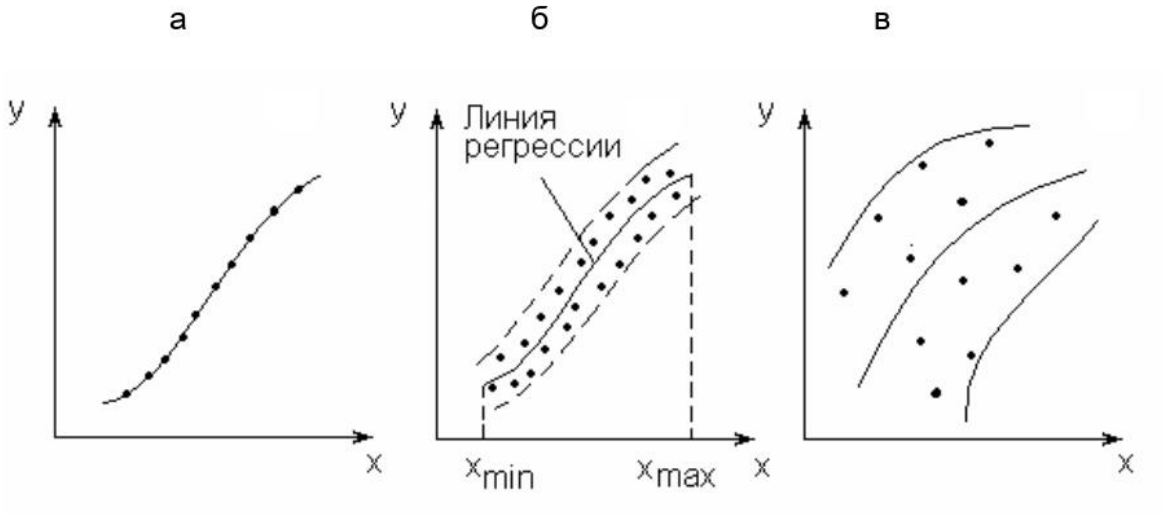

Рис.4.1. *Виды связей: а – функциональная связь, все точки лежат на линии; б – связь достаточно тесная, точки группируются возле линии регрессии, но не все они лежат на ней; в – связь слабая*

Стохастичность связи состоит в том, что одна случайная переменная y реагирует на изменение другой **X** изменением своего закона распределения (см. рис. 4.1, б). Таким образом, зависимая переменная принимает не одно конкретное значение, а некоторое из множества значений. Повторяя испытания, мы будем получать другие значения функции отклика, и одному и тому же значению X в различных реализациях будут соответствовать различные значения

y в интервале [x_{min}; x_{max}]. Искомая зависимость $\hat{y}$ =f(**X**) может быть найдена лишь в результате совместной обработки полученных значений **X** и y.

На рис. 4.1, б – это кривая зависимости, проходящая по центру полосы экспериментальных точек (математическому ожиданию), которые могут и не лежать на искомой кривой $\hat{y}$ =f(**X**), а занимают некоторую полосу вокруг нее. Эти отклонения вызваны погрешностями измерений, неполнотой модели и учитываемых факторов, случайным характером самих исследуемых процессов и другими причинами.

Анализ стохастических связей приводит 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, x_2, ..., x_i, ..., x_k, которые носят количественный характер;
3) задачи дисперсионного анализа – задачи, в которых переменные x_1, x_2, ..., x_i, ..., x_k имеют качественный характер, а исследуется и устанавливается степень их влияния на переменное y.

Стохастические зависимости характеризуются формой, теснотой связи и численными значениями коэффициентов уравнения регрессии.

Форма связи устанавливает вид функциональной зависимости $\hat{y}$=f(**X**) и характеризуется уравнением регрессии. Если уравнение связи линейное, то имеем линейную многомерную регрессию, в этом случае зависимость $\hat{y}$ от **X** описывается линейной зависимостью в k-мерном пространстве:

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_j, \qquad (4.2)$$

где b_0, ..., b_j, ..., b_k – коэффициенты уравнения. Для пояснения существа используемых методов ограничимся сначала случаем, когда x – скаляр. В общем

случае виды функциональных зависимостей в технике достаточно многообразны: показательные $y = b_0 x^{b_1}$, логарифмические $y = b_0 \lg(x)$ и т.д.

Заметим, что задача выбора вида функциональной зависимости – задача неформализуемая, так как одна и та же кривая на данном участке примерно с одинаковой точностью может быть описана самыми различными аналитическими выражениями. Отсюда следует важный практический вывод. Даже в наш век компьютеров принятие решения о выборе той или иной математической модели остается за исследователем. Только экспериментатор знает, для чего будет в дальнейшем использоваться эта модель, на основе каких понятий будут интерпретироваться ее параметры.

Крайне желательно при обработке результатов эксперимента вид функции $\hat{y}$=f(X) выбирать, исходя из условия ее соответствия физической природе изучаемых явлений или имеющимся представлениям об особенностях поведения исследуемой величины. К сожалению, такая возможность не всегда имеется, так как эксперименты чаще всего проводятся для исследования недостаточно или неполно изученных явлений.

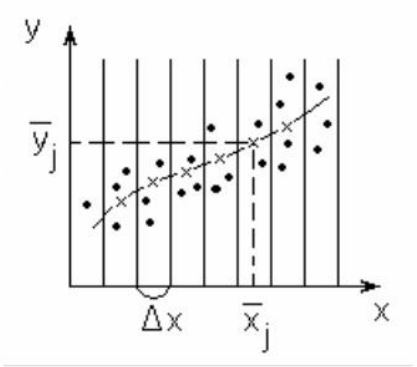

Рис.4.2. *К построению эмпирической линии регрессии*

При изучении зависимости $\hat{y}$=f(x) от одного фактора при заранее неизвестном виде функции отклика для приближенного определения вида уравнения регрессии полезно предварительно построить эмпирическую линию регрессии (рис.4.2). Для этого весь диапазон изменения x разбивают на равные интервалы Δx. Все точки, попавшие в данный интервал Δx_j, относят к его середине $\overline{x_j}$. Для этого подсчитывают частные средние для каждого интервала:

$$\overline{y_j} = \frac{\sum_{i=1}^{n_j} y_{ji}}{n_j}. \qquad (4.3)$$

Здесь n_j – число точек в интервале Δx_j, причем $\sum_{j=1}^{k^*} n_j = n$, где k* – число интервалов разбиения; n – объем выборки.

Затем последовательно соединяют точки $(\overline{x_j}; \overline{y_j})$ отрезками прямой. Полученная ломаная называется эмпирической линией регрессии. По виду эмпирической линии регрессии можно в первом приближении подобрать вид уравнения регрессии $\hat{y}$=f(x).

Под теснотой связи понимается степень близости стохастической зависимости к функциональной, т.е. показатель тесноты группирования экспериментальных данных относительно принятого уравнения модели (см. рис. 4.1,б,в). В дальнейшем уточним это положение.

4.2. Определение коэффициентов уравнения регрессии

Будем полагать, что вид уравнения регрессии уже выбран и требуется определить только конкретные численные значения коэффициентов этого уравнения **b**=$\{b_0,...,b_j,...,b_k\}$. Отметим предварительно, что если выбор вида уравнения регрессии, как это уже отмечалось, – процесс неформальный и не может быть полностью передан компьютеру, то расчет коэффициентов выбранного уравнения регрессии – операция достаточно формальная и ее следует решать с использованием компьютера. Это трудный и утомительный расчет, в котором человек не застрахован от ошибок, а компьютер выполнит его значительно быстрее и качественнее.

Существует два основных подхода к нахождению коэффициентов b_j. Выбор того или иного из них определяется целями и задачами, стоящими перед исследователем, точностью полученных результатов, их количеством и т.д.

Первый подход – интерполирование. Базируется на удовлетворении условию, чтобы функция $\hat{y}$**=(X,b)** совпадала с экспериментальными значениями в некоторых точках, выбранных в качестве опорных (основных, главных) y_i.

В этом случае для определения k+1 неизвестных значений параметров b_j используется система уравнений

$$f(x_i, b_0, ..., b_j, ..., b_k) = y_i, \quad 1 \le i \le n. \qquad (4.4)$$

В данном случае число независимых уравнений системы равно числу опорных точек, в пределе – n поставленных опытов. С другой стороны, для определения k+1 коэффициентов необходимо не менее k+1 независимых уравнений. Но если число n поставленных опытов и число независимых уравнений равно числу искомых коэффициентов k+1, то решение системы может быть единственно, а следовательно, точно соответствует случайным значениям исходных данных. Таким образом, в предельном случае, когда число коэффициентов уравнения регрессии равно числу экспериментальных точек n=k+1, все экспериментальные точки будут совпадать с их расчетными значениями. Следует заметить, что добиваться такого точного совпадения путем значительного увеличения числа коэффициентов уравнения регрессии часто просто неразумно, поскольку экспериментальные результаты получены с большей или меньшей погрешностью, и такая функция может просто не отражать действительного характера изменения исследуемой величины в силу влияния помех (возмущений) (рис.4.3).

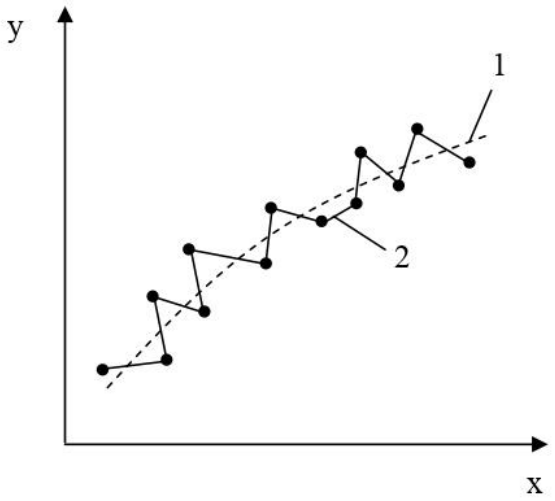

***Рис.4.3**. Аппроксимация функции с большим (1) и небольшим (2) числом коэффициентов b_i*

Таким образом, задача в конечном счете сводится к решению системы k+1 уравнений с k+1 неизвестными. Основная сложность такого решения связана с нелинейностью системы, хотя в принципе при использовании компьютера она преодолима.

При числе опытов n большем, чем k+1 искомых коэффициентов, число независимых уравнений системы избыточно. Избыточность информации можно использовать по-разному.

После определения численных значений k+1 параметров проверяется качество аппроксимации путем сопоставления значений функции и экспериментальных данных в оставшихся, неиспользованных точках. Если обнаруженные между ними расхождения превышают допустимые по условию точности, то процедуру определения коэффициентов b_j можно повторить, приняв в качестве опорных (основных) другие точки.

Таким образом, из этих уравнений в разных комбинациях можно составить несколько систем уравнений, каждая из которых в отдельности даст свое решение. Но между собой они будут несовместимыми. Каждое решение будет соответствовать своим значениям коэффициентов b_j. Если все их построить на графике, то получим целый пучок аппроксимирующих кривых.

Это открывает при n>k+1 совершенно новые возможности. Во-первых, этот пучок кривых показывает форму и ширину области неопределенности проведенного эксперимента. Во-вторых, может быть произведено усреднение всех найденных кривых и полученная усредненная кривая будет гораздо точнее и достовернее описывать исследуемое явление, так как она в значительной степени освобождена от случайных погрешностей, приводивших к разбросу отдельных экспериментальных точек. Поясним суть этого подхода на примере двух методов.

1. Метод избранных точек (рис. 4.4). На основании анализа данных выдвигают гипотезу о виде (форме) зависимости f(**X**). Предположим, что она линейная, т.е. статистическая связь – это линейная одномерная регрессия

$$\hat{y} = b_0 + b_1 x. \qquad (4.5)$$

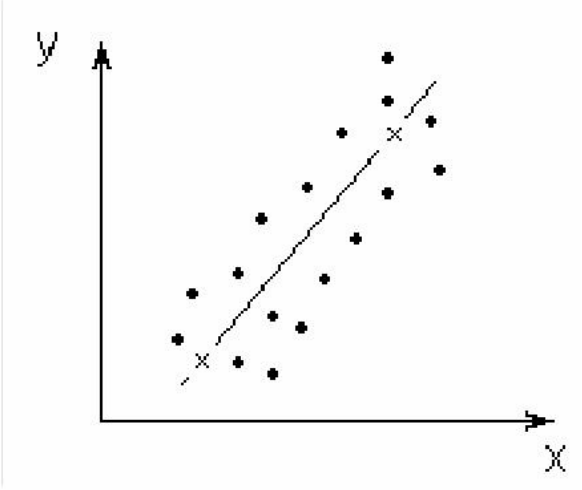

***Рис.4.4**. Метод избранных точек:*
× – избранные точки

Выбирают две наиболее характерные, по мнению исследователя, точки, через которые и проходит линия регрессии (рис. 4.4). Задача вычисления коэффициентов b_0 и b_1 в этом случае тривиальна. Если предполагается, что уравнение регрессии более высокого порядка, то соответственно увеличивают число избранных точек. Недостатки такого подхода очевидны, так как избранные точки выбираются субъективно, а подавляющая часть экспериментального материала не используется для определения параметров (коэффициентов) уравнения регрессии, хотя ее можно использовать в дальнейшем для оценки надежности полученного уравнения.

2. Метод медианных центров. Сущность этого метода поясняет рис.4.5. Обведенное контуром поле точек делят на несколько частей, число которых равно числу определяемых коэффициентов уравнения регрессии. В каждой из этих частей находят медианный центр, т.е. пересечение вертикали и горизонтали слева и справа, выше и ниже которых оказывается равное число точек. Затем через эти медианные центры проводят плавную кривую и из решения системы уравнений определяют коэффициенты регрессии b_j. Так, в случае линейной зависимости (4.5) поле делится на две группы. Определяют средние значения $\overline{x_I}, \overline{y_I}; \overline{x_{II}}, \overline{y_{II}}$ для каждой из групп, а неизвестные коэффициенты b_0, b_1 определяют из решения системы уравнений:

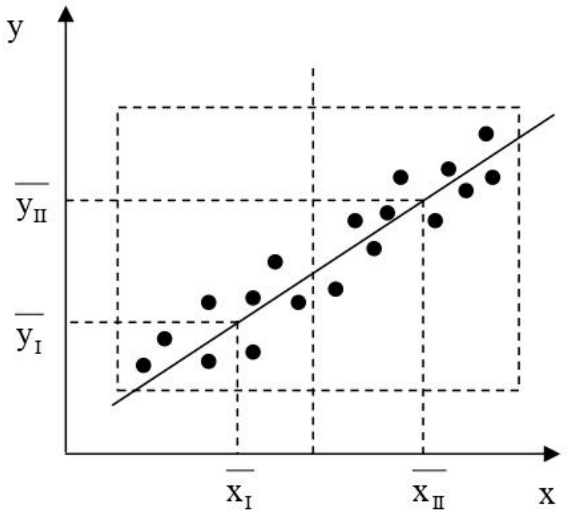

***Рис.4.5.** Метод медианных точек*

$$\overline{y_I} = b_0 + b_1 \overline{x_I};$$
$$\overline{y_{II}} = b_0 + b_1 \overline{x_{II}}. \qquad (4.5а)$$

Если при выборе вида уравнения регрессии число его коэффициентов b_j окажется больше числа уравнений (имеющихся результатов измерений) k+1>n, система (4.4) не будет иметь однозначного решения. В этом случае необходимо либо уменьшить число определяемых коэффициентов k+1, либо увеличить число опытов n.

Второй подход – метод наименьших квадратов. Усреднение несовместимых решений избыточной системы уравнений n>k+1 может быть преодолено методом наименьших квадратов, который был разработан еще Лежандром и Гауссом. Таким образом, метод наименьших квадратов – это «новинка» почти 200-летней давности. Сегодня, благодаря возможностям компьютеров, этот метод вступил, по существу, в полосу своего «ренессанса».

Определение коэффициентов b_j методом наименьших квадратов основано на выполнении требования, чтобы сумма квадратов отклонений экспериментальных точек от соответствующих значений уравнения регрессии была минимальна. Заметим, что, в принципе, можно оперировать и суммой других четных степеней этих отклонений, но тогда вычисления будут сложнее. Однако руководствоваться суммой отклонений нельзя, так как она может оказаться малой при больших отклонениях отрицательного знака.

Математическая запись приведенного выше требования имеет вид

$$\Phi(b_0, b_1, ..., b_j, ..., b_k) = \sum_{i=1}^{n} [f(x_i, b_0, b_1, ..., b_j, ..., b_k) - y_i]^2 \to \min_{b_j}, \qquad (4.6)$$

где n – число экспериментальных точек в рассматриваемом интервале изменения аргумента x.

Необходимым условием минимума функции $\Phi(b_0,b_1,...,b_j,...,b_k)$ является выполнение равенства

$$\partial \Phi / \partial b_j = 0, \quad 0 \le j \le k \qquad (4.7)$$

или

$$\sum_{i=1}^{n}[f(x_i,b_0,b_1,...,b_j,...,b_k)-y_i]\frac{\partial f(x_i)}{\partial b_j}=0,\quad 0\le j\le k. \qquad (4.7а)$$

После преобразований получим

$$\sum_{i=1}^{n}[f(x_i,b_0,b_1,...,b_j,...,b_k)\frac{\partial f(x_i)}{\partial b_j}-\sum_{i=1}^{n}y_i\frac{\partial f(x_i)}{\partial b_j}=0. \qquad (4.8)$$

Система уравнений (4.8) содержит столько же уравнений, сколько неизвестных коэффициентов b_0, b_1,..., b_k входит в уравнение регрессии, и называется в математической статистике системой нормальных уравнений.

Поскольку Ф≥0 при любых b_0, ..., b_k, величина Ф обязательно должна иметь хотя бы один минимум. Поэтому если система нормальных уравнений имеет единственное решение, оно и является минимумом для этой величины.

Расчет регрессионных коэффициентов методом наименьших квадратов можно применять при любых статистических данных, распределенных по любому закону.

4.3. Определение тесноты связи между случайными величинами

Определив уравнение теоретической линии регрессии, необходимо дать количественную оценку тесноты связи между двумя рядами наблюдений. Линии регрессии, проведенные на рис. 4.1, б, в, одинаковы, однако на рис. 4.1, б точки значительно ближе (теснее) расположены к линии регрессии, чем на рис. 4.1, в.

При корреляционном анализе предполагается, что факторы и отклики носят случайный характер и подчиняются нормальному закону распределения.

Тесноту связи между случайными величинами характеризуют корреляционным отношением ρ_{xy}. Остановимся подробнее на физическом смысле данного показателя. Для этого введем новые понятия.

Остаточная дисперсия $S^2_{y\,ост}$ характеризует разброс экспериментально наблюдаемых точек относительно линии регрессии и представляет собой показатель ошибки предсказания параметра y по уравнению регрессии (рис. 4.6):

$$S^2_{y\,ост} = \frac{1}{n-l}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2 = \frac{1}{n-1-k}\sum_{i=1}^{n}[y_i - f(x_i, b_0, b_1, ..., b_k)]^2, \qquad (4.9)$$

где $l=k+1$ – число коэффициентов уравнения модели.

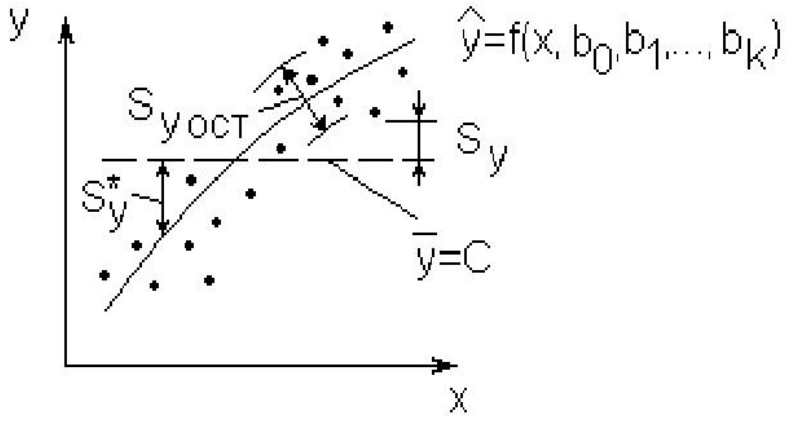

Рис.4.6. *К определению дисперсий*

Общая дисперсия (дисперсия выходного параметра) S^2_y характеризует разброс экспериментально наблюдаемых точек относительно среднего значения $\bar{y}$, т.е. линии С (см. рис. 4.6):

$$S^2_y = \frac{1}{n-1}\sum_{i=1}^{n}[y_i - \bar{y}]^2, \qquad (4.10)$$

где $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Средний квадрат отклонения линии регрессии от среднего значения линии $\bar{y} = C$ (см. рис. 4.6):

$$S^{*2}_y = \frac{1}{n-1}\sum_{i=1}^{n}[\hat{y}_i - \bar{y}]^2 = \frac{1}{n-1}\sum_{i=1}^{n}[f(x_i, b_0, b_1, ..., b_k) - \bar{y}]^2. \qquad (4.11)$$

Очевидно, что общая дисперсия S^2_y (сумма квадратов относительно среднего значения $\bar{y}$) равна остаточной дисперсии $S^2_{y\,ост}$ (сумме квадратов относительно линии регрессии) плюс средний квадрат отклонения линии регрессии S_y^{*2} (сумма квадратов, обусловленная регрессией).

$$S_y^2 = S_{y\,ост}^2 + S_y^{*2}. \qquad (4.11а)$$

Разброс экспериментально наблюдаемых точек относительно линии регрессии характеризуется безразмерной величиной – выборочным корреляционным отношением, которое определяет долю, которую привносит величина X в общую изменчивость случайной величины Y.

$$\rho_{xy}^{*} = \sqrt{\frac{S_y^2 - S_{y\,ост}^2}{S_y^2}} = \sqrt{\frac{S_y^{*2}}{S_y^2}} = \frac{S_y^{*}}{S_y}. \qquad (4.12)$$

Проанализируем свойства этого показателя.

1. В том случае, когда связь является не стохастической, а функциональной, корреляционное отношение равно 1, так как все точки корреляционного поля оказываются на линии регрессии, остаточная дисперсия равна $S_{y\,ост}^2 = 0$, а $S_y^{*2} = S_y^2$ (рис. 4.7, а).

2. Равенство нулю корреляционного отношения указывает на отсутствие какой-либо тесноты связи между величинами x и y для данного уравнения регрессии, поскольку разброс экспериментальных точек относительно среднего значения и линии регрессии одинаков, т.е. $S_y^2 = S_{y\,ост}^2$ (рис. 4.7, б).

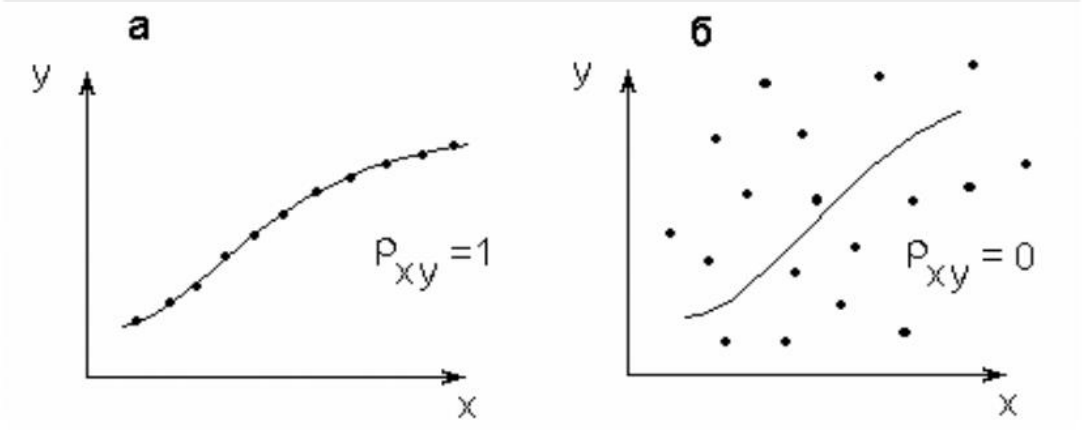

Рис. 4.7. *Значения выборочного корреляционного отношения* ρ_{xy}*:*

а – функциональная связь; б – отсутствие связи

3. Чем ближе расположены экспериментальные данные к линии регрессии, тем теснее связь, тем меньше остаточная дисперсия и тем больше корреляционное отношение.

Следовательно, корреляционное отношение может изменяться в пределах от 0 до 1.