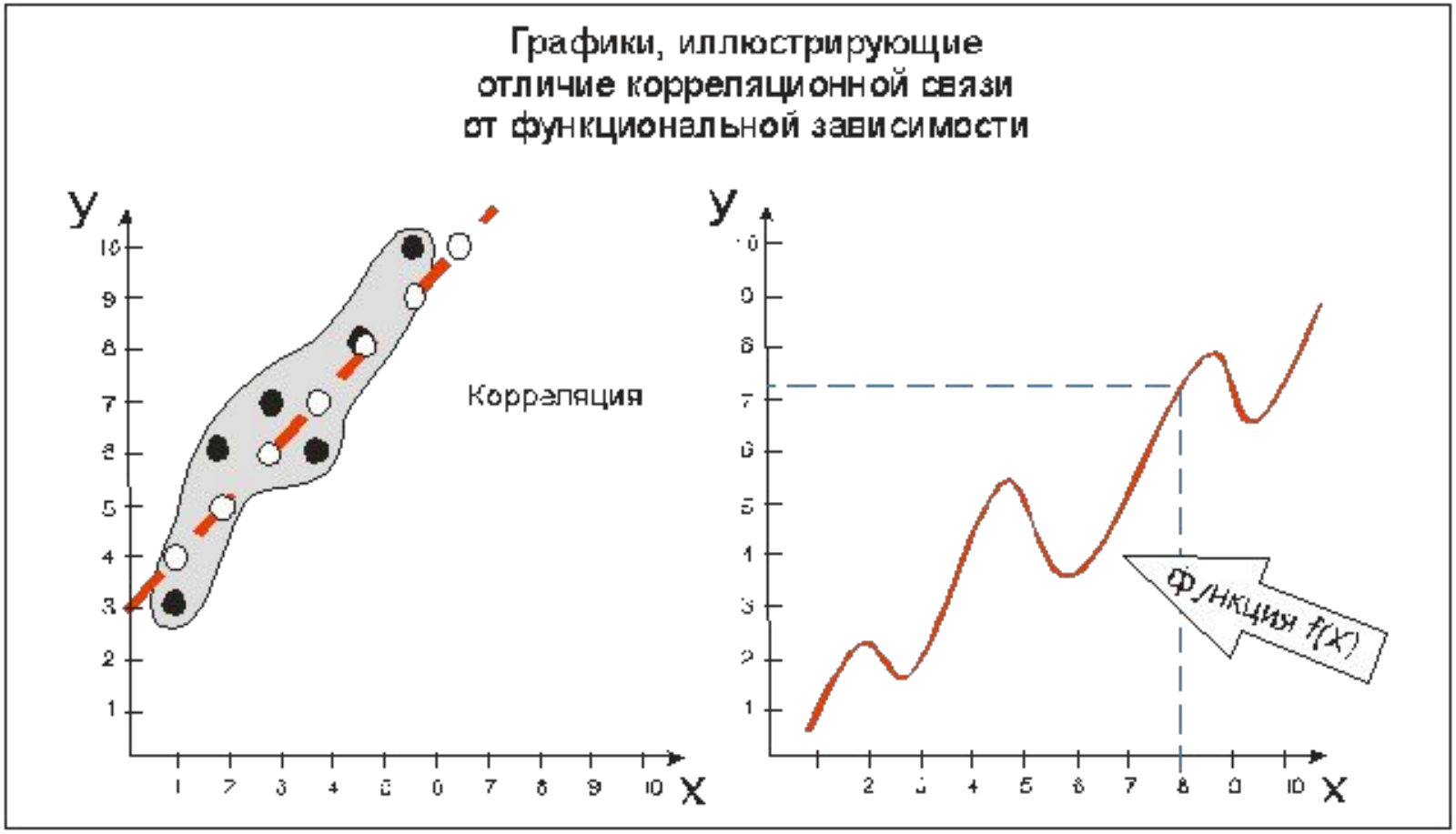
КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ



Функция, во-первых, непрерывна, тогда как при корреляционной зависимости значения, принимаемые признаком, дискретны. Во-вторых, функциональная зависимость предполагает взаимно однозначное соответствие аргумента ***x*** и функции ***f(x)***, вероятностная же зависимость допускает некий условный диапазон, в который предположительно (с такой-то долей вероятности) попадает значение признака **yi** при значении **xi** признака **x**.

1. Кустистость растений *(x):* 4; 6; 10; 12, в среднем 8.

Вес растений в *г (y):* 30; 34; 42; 46, в среднем 38.

Вес растений (*y*)	Кустистость (*x*)			
	4	6	10	12
30	1			
34		1		
42			1	
46				1

2. Кустистость растений *(x):* 4; 6; 10; 12, в среднем 8.
Вес растений в *г (y):* 46; 42; 34; 30, в среднем 38.

Вес растений (y)	Кустистость (x)			
	4	6	10	12
30				1
34			1	
42		1		
46	1			

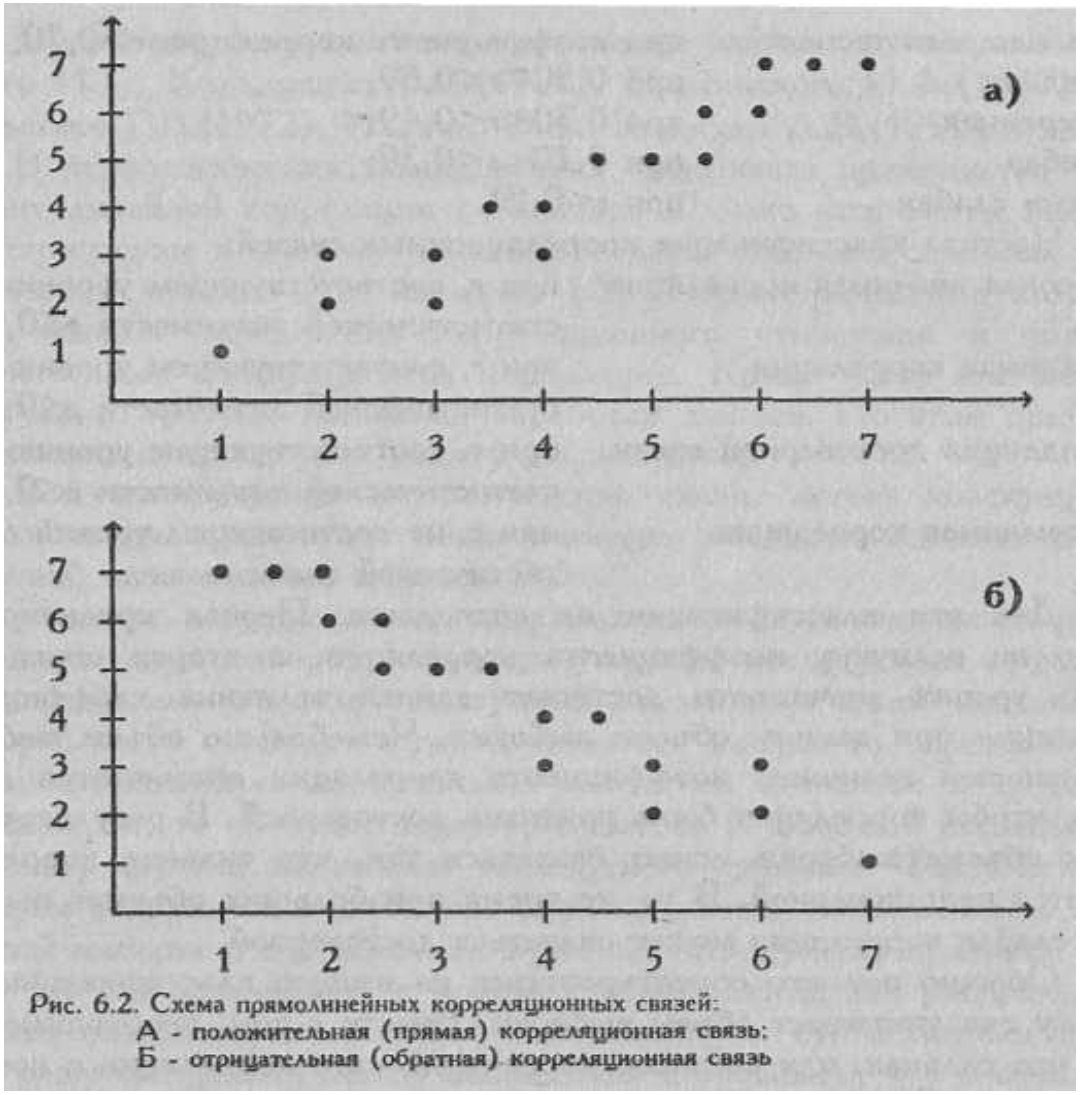


Рис. 6.2. Схема прямолинейных корреляционных связей:
А - положительная (прямая) корреляционная связь;
Б - отрицательная (обратная) корреляционная связь

3. Кустистость растений *(x):* 4; 6; 10; 12, в среднем 8.
Вес растений в *г (y):* 42; 30; 46; 34, в среднем 38

Вес растений (*y*)	Кустистость (*x*)			
	4	6	10	12
30		1		
34				1
42	1			
46			1	

Ранговый коэффициент корреляции Спирмена *(rs)*

$$r_s = 1 - \frac{6 \cdot \Sigma (x - y)^2}{n \cdot (n^2 - 1)},$$

где *x* и *y* — ранги по каждому признаку; *n* — число членов в совокупности. Формула может быть упрощена, если выражение *(x*—y/)2 заменить на *D2.* Тогда

$$r_s = 1 - \frac{6 \cdot \Sigma D^2}{n \cdot (n^2 - 1)}.$$

Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами.
Пусть даны две выборки $x^m = (x_1, \cdots, x_m),\ y^m = (y_1, \cdots, y_m)$;

коэффициент корреляции Пирсона рассчитывается по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{m} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \overline{x})^2 \sum_{i=1}^{m} (y_i - \overline{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}},$$

Сильная	или	тесная **более 0,70**
Средняя		от 0,50 до 0,69
Умеренная		от 0,30 до 0,49
Слабая		от 0,20 до 0,29
Очень слабая		меньше 0,19

Номер испытуемого	Результат I теста	Результат II теста	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$\frac{(x_i - \bar{x})}{(y_i - \bar{y})}$
1	14	21	−7,3	53,3	−0,1	0,001	0,7
2	30	22	8,7	75,7	0,9	0,8	7,8
3	16	18	−5,3	28,1	−3,1	9,6	16,4
4	18	20	−3,3	10,9	−1,1	1,2	3,6
5	25	24	3,7	13,7	2,9	8,4	10,7
6	17	19	−4,3	18,5	−2,1	4,4	9,0
7	21	23	−0,3	0,1	1,9	3,6	−0,6
8	29	23	7,7	59,3	1,9	3,6	14,6
9	24	22	2,7	7,3	0,9	0,8	2,4
10	19	19	−2,3	5,3	−2,1	4,4	4,8
Σ	213	211	0	272,2	0	36,8	69,4
$\bar{x}$	21,3	21,1	—	—	—	—	—

$$\sigma_x = \sqrt{\frac{(x_i - \bar{x})^2}{(n-1)}} = \sqrt{\frac{272,2}{10}} = 5,22; \qquad \sigma_y = \sqrt{\frac{(y_i - \bar{y})^2}{(n-1)}} = \sqrt{\frac{36,8}{10}} = 1,92;$$

$$r_{xy} = \frac{\Sigma[(x_i - \bar{x})(y_i - \bar{y})]}{(n-1)\sigma_x\sigma_y} = \frac{69,4}{10 \cdot 5,22 \cdot 1,92} = 0,69.$$

y

Статистическая проверка наличия корреляции

- **Гипотеза Но:** : отсутствует линейная связь между выборками x и y ($r_{xy} = 0$)
- **Статистика критерия:**

$$T = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \sim t_{n-2}$$

– распределение Стьюдента с степенями свободы.

Графическое представление корреляции

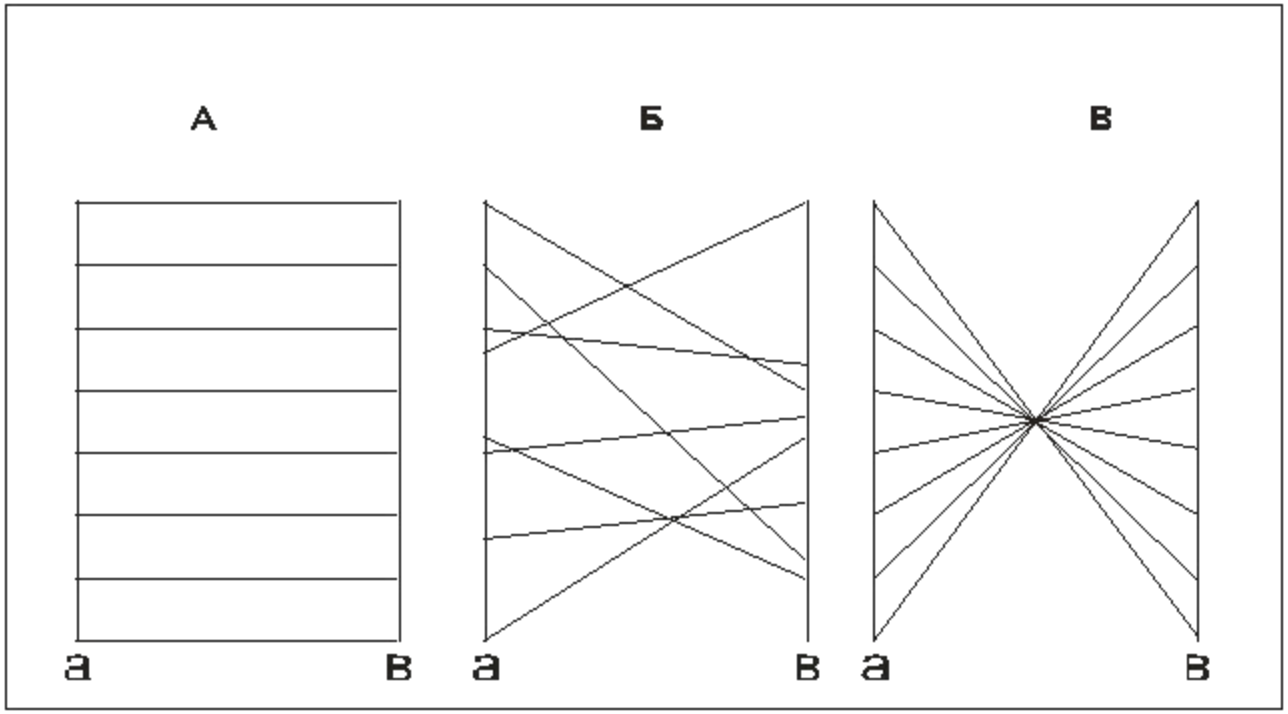


Рис. А Показана жесткая связь с коэффициентом корреляции, равным +1. Увеличению признака А сопутствует увеличение признака В на ту же величину.

Рис. Б Нет взаимосвязи между изменениями А и В. При увеличении А, В может меняться как в сторону увеличения, так и в сторону уменьшения.

Рис. В Пример сильной корреляции с коэффициентом -1. Увеличение признака А сопровождается пропорциональным уменьшением признака В.

Линейная корреляция

- Предположим, что мы располагаем выборкой данных о какой-то группе объектов.
- Пусть эти объекты обладают общими родовыми особенностями (примерно одинаковы).
- Пусть, к тому же, у каждого из объектов можно количественно измерить, как минимум, два каких-либо параметра.

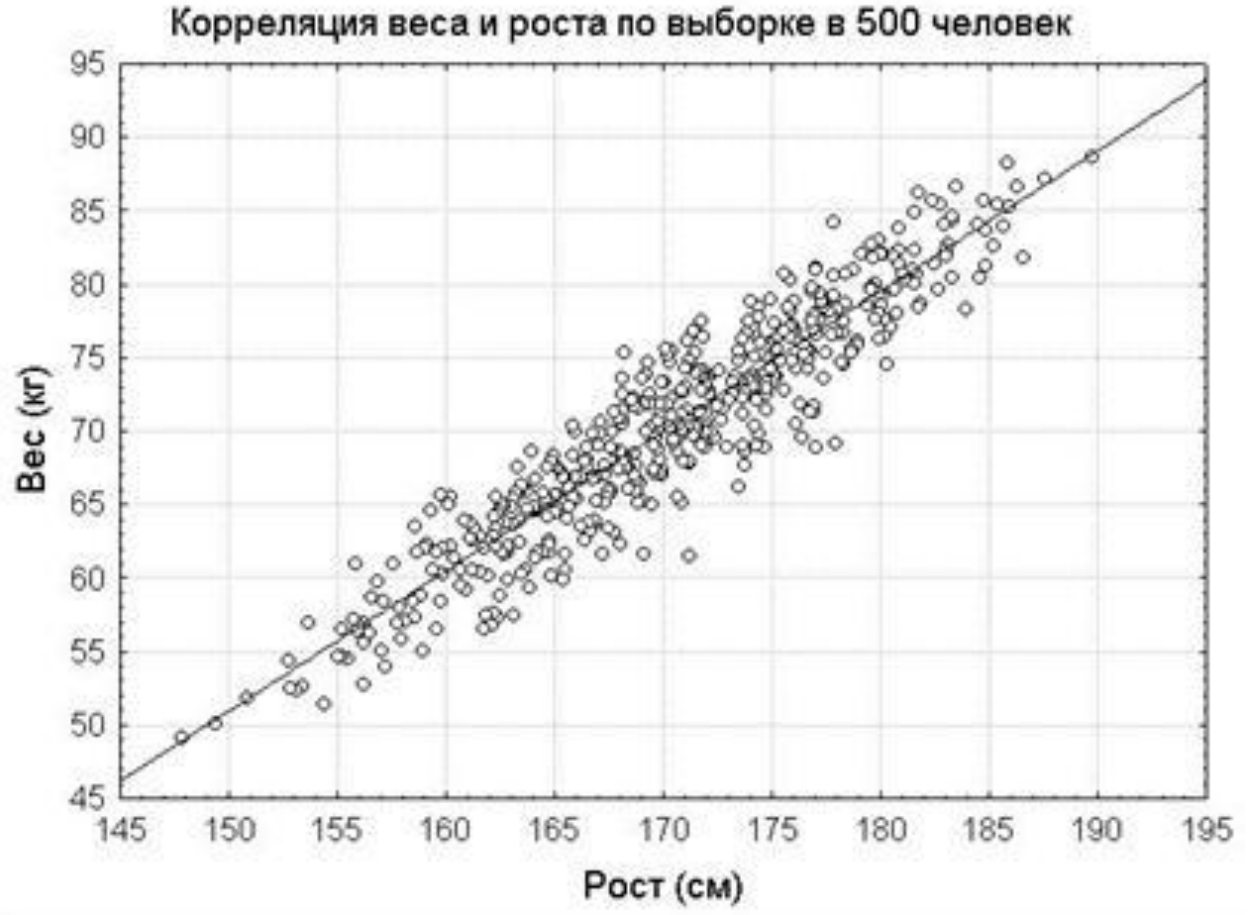
При этих обстоятельствах открывается возможность для подсчета линейной корреляции между двумя (или более) признаками, присущими этим объектам.

Например, такими выборками данных могут служить сведения о:

- группе людей, рост и вес тела которых мы измеряем;

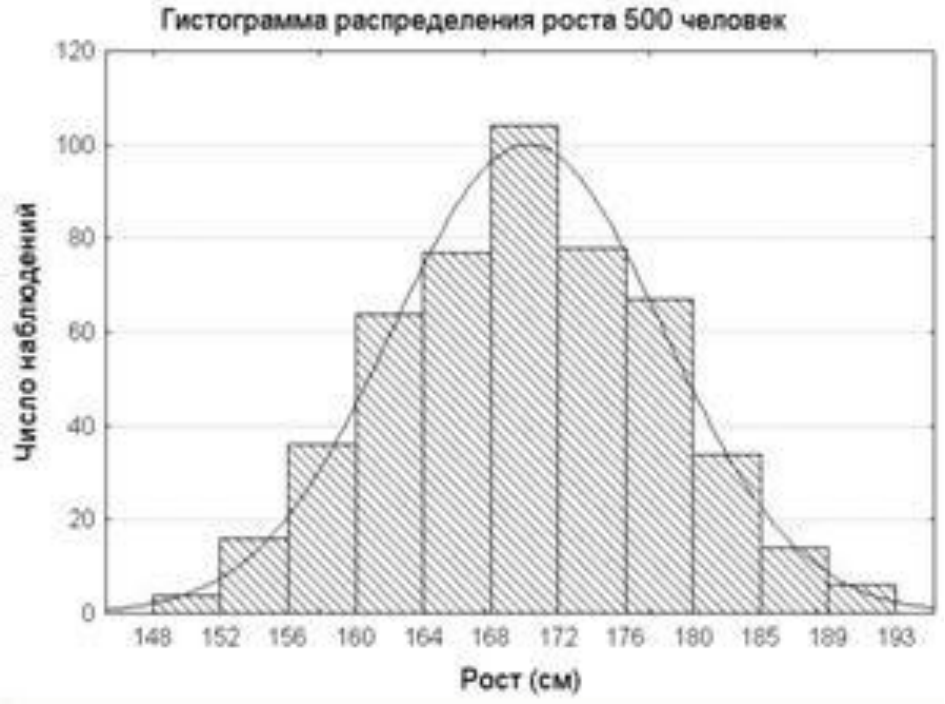
- длине и ширине лепестка какого-нибудь цветка.

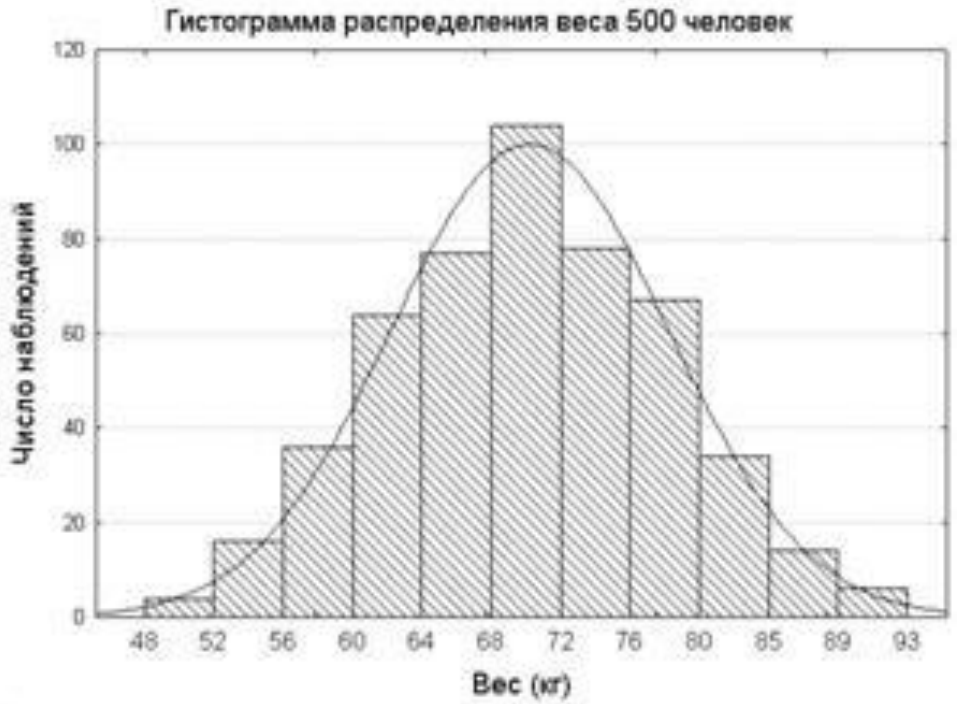
Двумерная диаграмма рассеяния, отражающая линейную корреляцию между ростом и весом человека.



Вычисление же коэффициентов корреляции Пирсона предполагает, что каждый из анализируемых количественных признаков, подчиняется нормальному закону.

Гистограммы распределения для роста и веса.





Коэффициент корреляции Пирсона как количественная мера связи

Коэффициент корреляции возведенный в квадрат (r^2) называется коэффициентом детерминации и отражает долю вариативности одной переменной, которая может быть предсказана на основе другой. Можно сказать, что коэффициент детерминации характеризует долю общих факторов определяющих поведение обеих переменных. Если одна из переменных является независимой, а другая зависимой, то r^2 представляет собой долю дисперсии (или вариативности) зависимой переменной объясняемой влиянием независимой переменной.

Так, если коэффициент корреляции между переменными равен 0,7, то 49% (т.е. $0,7^2$) вариативности одной переменной можно предсказать на основе знания другой. Остаток вариативности (в данном случае 51%) обусловлен другими переменными или случайной ошибкой.

Регрессия

- •Моделирование, описание зависимости между переменными

•Количественная оценка поведения отклика при изменении предиктора **->>** *уравнение регрессии*

•Предсказание значений переменной отклика при заданных значениях предиктора **->>** *прогноз*

- Функция *f*(*x*2 , *x*3 , …, *xm*), описывающая зависимость показателя от параметров, называется уравнением (функцией) регрессии.

- Требуется: установить количественную взаимосвязь между показателем и факторами. В таком случае задача регрессионного анализа понимается как задача выявления такой функциональной зависимости *y* = f(x2 , x3 , …, xm*), которая наилучшим образом описывает имеющиеся экспериментальные данные.

- Регрессия — зависимость математического ожидания (например, среднего значения) случайной величины от одной или нескольких других случайных величин (свободных переменных), то есть $E(y|x) = f(x)$
- Регрессионным анализом называется поиск такой функции , которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих. $y = f(x) + \nu,$
- где — f функция регрессионной зависимости, а — ν аддитивная случайная величина с нулевым матожиданием.

$$y(x_1, x_2, ..., x_p) = E(Y/(\mathrm{X}_1 = \mathrm{x}_1, \mathrm{X}_2 = \mathrm{x}_2, ..., \mathrm{X}_\mathrm{p} = \mathrm{x}_\mathrm{p}))$$

Уравнение регрессии

$Y = b0 + b1X$

$y = a + bx$

- Y –зависимая переменная, отклик

- X –независимая переменная, предиктор, фактор

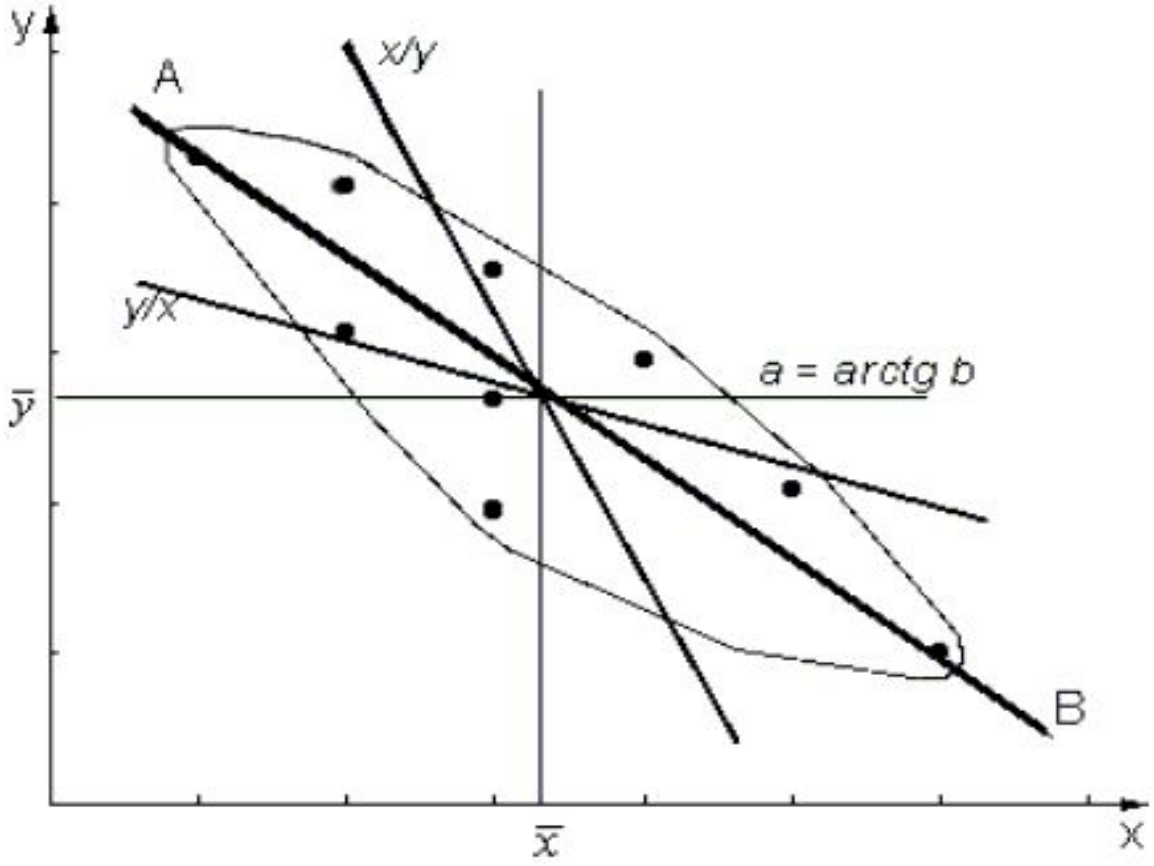
- b0 ,a–ожидаемое значение Y при X= 0

свободный член; графически он представляет отрезок ординаты (y) в системе прямоугольных координат.

- b1– **коэффициент регрессии**

угол наклона графика по отношению к оси X, среднее изменение Y на единицу изменения X в выборке

Схема линий регрессии Y по X и X по Y в системе прямоугольных координат



Коэффициенты уравнения парной линейной регрессии

- Y = a1 + by/xX — прямое
- и X = a2 + bx/yY — обратное, (2.2)
- где: a и b – коэффициенты, или параметры, которые надлежит определить.
- Значение коэффициентов регрессии вычисляется по формуле:

$$b_{x/y} = r \cdot \frac{\sigma_x}{\sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2}$$

$$b_{y/x} = r \cdot \frac{\sigma_y}{\sigma_x} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Коэффициенты a определяются по формуле

$$a_1 = \bar{y} - b_{y/x} \cdot \bar{x}$$

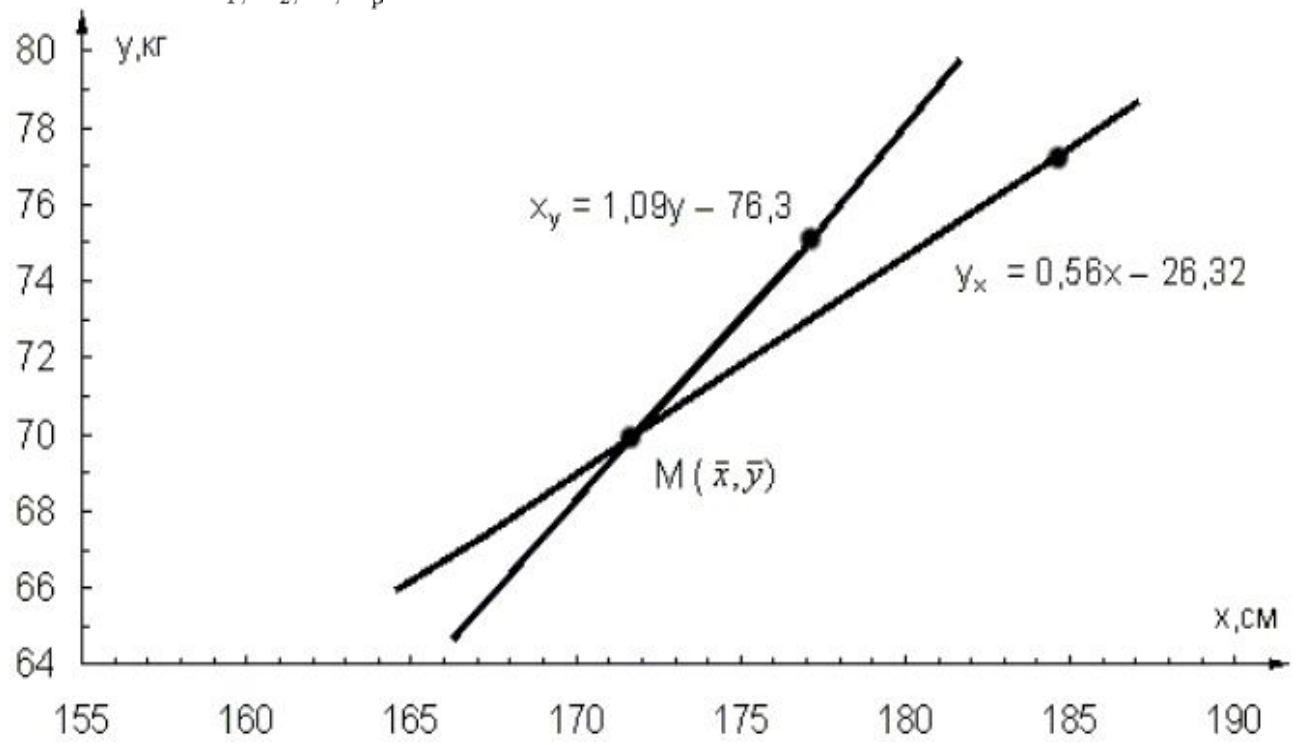
$$a_2 = \bar{x} - b_{x/y} \cdot \bar{y}$$

Способ наименьших квадратов

В основу этого способа положена теорема, согласно которой сумма квадратов отклонений вариант (xi) от средней арифметической () есть величина наименьшая, т.е.

$$\sum_{k=1}^{N}(Y_k - \hat{Y}_k)^2 \to min$$

то функция $y(x_1, x_2, ..., x_p)$ называется регрессией величины Y по величинам $X_1, X_2, ..., X_p$



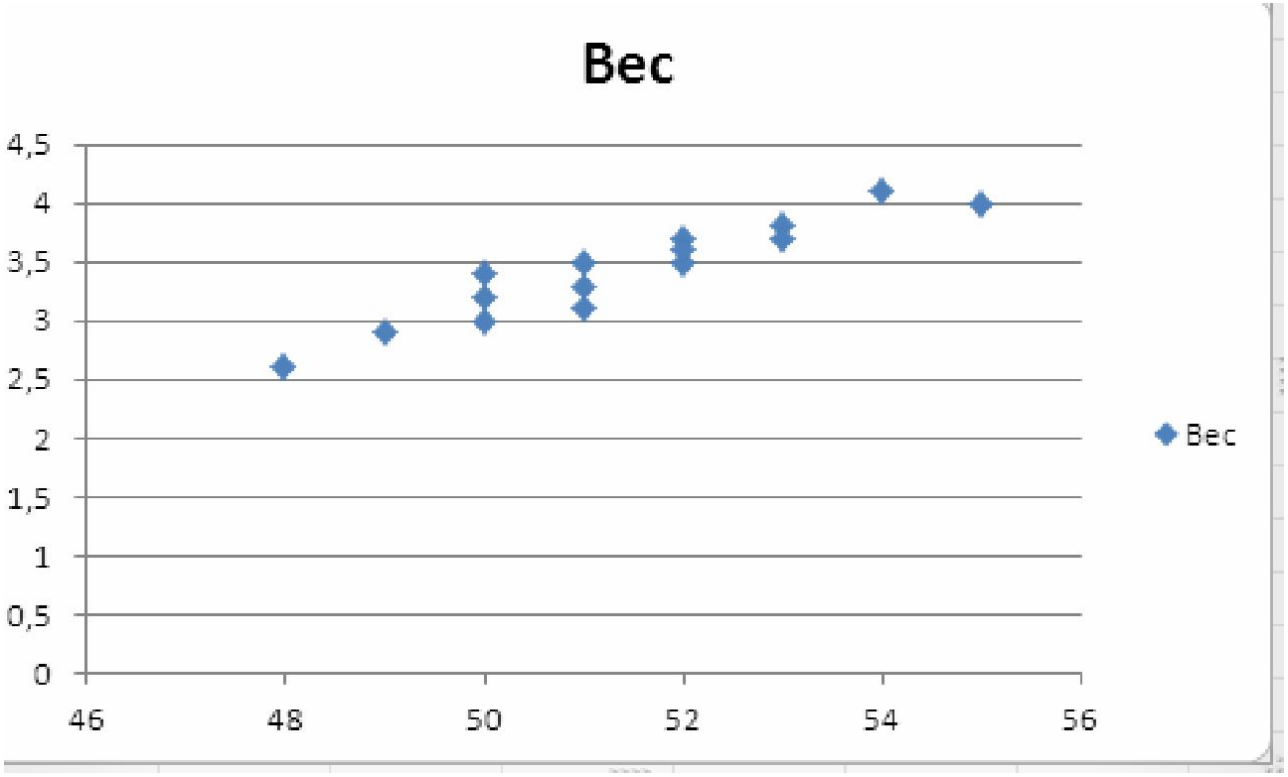
Графическое изображение эмпирического уравнения регрессии.

•Имеются данные измерений роста X (см) и веса Y (кг) новорождённых:

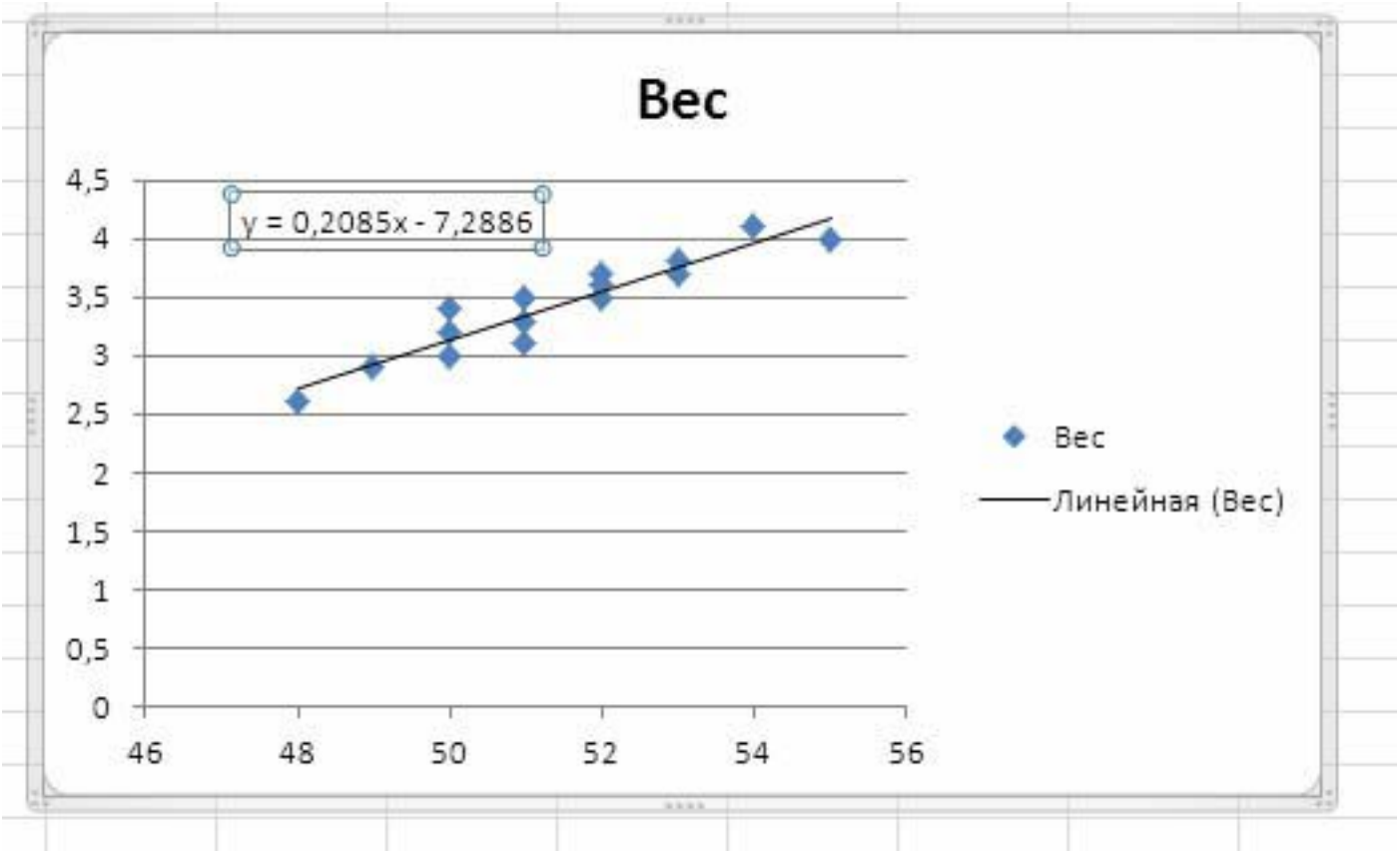
	A	B	C	D
1	Рост	Вес		
2	50	3,2		
3	49	2,9		
4	51	3,3		
5	48	2,6		
6	51	3,1		
7	52	3,5		
8	50	3		
9	52	3,7		
10	53	3,8		
11	50	3,4		
12	54	4,1		
13	51	3,5		
14	55	4		
15	53	3,7		
16	52	3,6		

Проведите регрессионный анализ: составьте уравнение линейной регрессии и таблицу наилучшего соответствия веса для роста: 50, 51 и 52 см.
Оцените вес ребенка ростом 55 см.

Корреляционное поле лучше всего описывается линейным уравнением



Линия регрессии



Расчет наилучшего соответствия веса для роста: 50, 51 и 52 см, используя уравнение регрессии $y = 0{,}2085x - 7{,}2886$.

	A	B
10	53	3,8
11	50	3,4
12	54	4,1
13	51	3,5
14	55	4
15	53	3,7
16	52	3,6
17		
18	50	3,1364
19	51	3,3449
20	52	3,5534

Оценка веса ребенка ростом 55 см. Используем уравнение линейной регрессии.

14	55	4	
17			Расчет
18	50	3,1364	
19	51	3,3449	
20	52	3,5534	
21			Прогноз
22	55	4,1789	
23			