

Лекция 8. Понятие корреляции, диаграмма рассеивания. Простейшие разновидности корреляции. Коэффициент корреляции. Величина корреляции и сила связи

Корреляционная связь – это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого.

Корреляционные связи различаются по **форме, направлению и степени** (силе).

По форме корреляционная связь может быть прямолинейной или криволинейной. Прямолинейной может быть, например, связь между количеством тренировок на тренажере и количеством правильно решаемых задач в контрольной сессии. Криволинейной может быть, например, связь между уровнем мотивации и эффективностью выполнения задачи. При повышении мотивации эффективность выполнения задачи сначала возрастает, затем достигается оптимальный уровень мотивации, которому соответствует максимальная эффективность выполнения задачи; дальнейшему повышению мотивации сопутствует уже снижение эффективности.

По направлению корреляционная связь может быть положительной (прямой) и отрицательной (обратной). При положительной прямолинейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого. При отрицательной корреляции соотношения обратные. При положительной корреляции коэффициент корреляции имеет положительный знак, например $r_{xy} = 0,207$, при отрицательной корреляции – отрицательный знак, например, $r_{xy} = -0,207$. **Степень, сила или теснота** корреляционной связи определяется по величине коэффициента корреляции. Сила связи не зависит от её направленности и определяется по абсолютному значению коэффициента корреляции. Максимальное возможное абсолютное значение коэффициента корреляции $r_{xy} = 1$; минимальное $r_{xy} = 0$.

Рассмотрим общую классификация корреляционных связей:

сильная, или тесная при коэффициенте корреляции $|r_{xy}| > 0,70$;

средняя при $0,50 < |r_{xy}| < 0,69$;

умеренная при $0,30 < |r_{xy}| < 0,49$;

слабая при $0,20 < |r_{xy}| < 0,29$;

очень слабая при $|r_{xy}| < 0,19$

Корреляционная связь — такая связь, при которой каждому определенному значению одного признака соответствует несколько значений другого взаимосвязанного с ним признака (связь между ростом и массой тела человека; связь между температурой тела и частотой пульса и

др.). Корреляционная связь характерна для медико-биологических процессов.

Величина, характеризующая направление и силу связи между признаками называется коэффициентом корреляции, он одним числом дает представление о направлении и силе связи между признаками (явлениями), пределы его колебаний от 0 до ± 1

· **Методика и порядок вычисления коэффициента корреляции**

1) Метод квадратов

построить вариационные ряды для каждого из сопоставляемых признаков, обозначив первый и второй ряд чисел соответственно x и y ; определить для каждого вариационного ряда средние значения (M_1 и M_2);

найти отклонения (d_x и d_y) каждого числового значения от среднего значения своего вариационного ряда;

полученные отклонения перемножить ($d_x \times d_y$)

каждое отклонение возвести в квадрат и суммировать по каждому ряду ($\sum d_x^2$ и $\sum d_y^2$)

подставить полученные значения в формулу расчета коэффициента корреляции:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{\sum (d_x \times d_y)}{\sqrt{(\sum d_x^2 \times \sum d_y^2)}}$$

Практическое значение установления корреляционной связи.

Выявление причинно-следственной между факторными и результативными признаками (при оценке физического развития, для определения связи между условиями труда, быта и состоянием здоровья, при определении зависимости частоты случаев болезни от возраста, стажа, наличия производственных вредностей и др.)

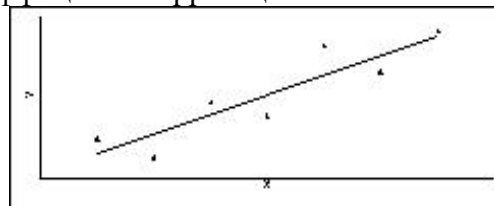
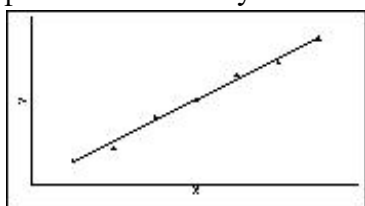
Зависимость параллельных изменений нескольких признаков от какой-то третьей величины. Например, под воздействием высокой температуры в цехе происходят изменения кровяного давления, вязкости крови, частоты пульса и др.

Способы представления корреляционной связи

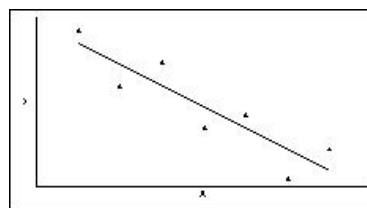
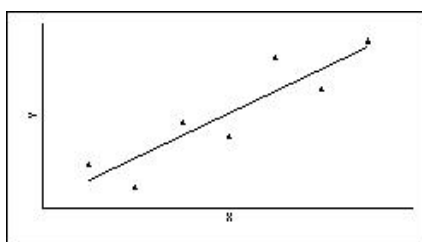
1) График (диаграмма рассеяния). Простейшим визуальным способом выявить наличие взаимосвязи между количественными переменными является построение **диаграммы рассеяния** (*scatterplot*). Это график, на котором по горизонтальной оси (X) откладывается одна переменная, по вертикальной (Y) другая. Каждому объекту на диаграмме соответствует точка, координаты которой равняются значениям пары выбранных для анализа переменных.

Визуально о силе связи можно судить по тому, насколько тесно расположены точки-объекты около линии регрессии. Чем ближе точки к линии регрессии, тем сильнее связь.

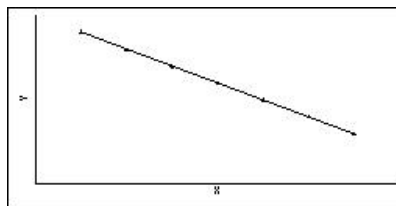
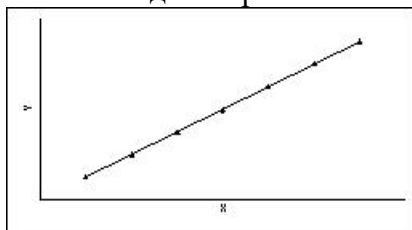
На каком из двух графиков связь между признаками сильнее (теснее), т.е. какому из графиков соответствует более высокий коэффициент корреляции?



Коэффициент парной корреляции r принимает значения в диапазоне от -1 до $+1$. Положительные значения коэффициента корреляции r свидетельствуют о положительной связи между признаками, отрицательные – об отрицательной связи.

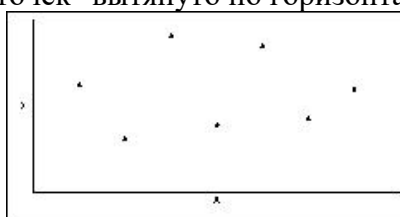


Если $r = 1$, то между двумя переменными существует **функциональная положительная линейная связь**, т.е. на диаграмме рассеяния соответствующие точки лежат на одной прямой с положительным наклоном.



Если $r = -1$, то между двумя переменными существует **функциональная отрицательная линейная зависимость**, т.е. на диаграмме рассеяния соответствующие точки лежат на одной прямой с отрицательным наклоном.

Если $r = 0$, то рассматриваемые переменные **линейно независимы**, т.е. на диаграмме рассеяния облако точек "вытянуто по горизонтали".



коэффициент корреляции показывает силу связи между признаками
Знак коэффициента говорит о направлении связи

Построение диаграммы рассеяния

Диаграмму строят следующим образом. В первую очередь необходимо собрать пары данных (x, y) , между которыми будет исследоваться зависимость (не менее 30 пар данных). Находим максимальные и минимальные значения и для x , и для y . Выбираем шкалы на горизонтальной и вертикальной оси так, чтобы обе длины рабочих частей получились примерно одинаковыми, тогда диаграмма будет легче читаться. Желательно на каждой оси взять от 3 до 10 градаций. Если одна переменная – фактор, а вторая – характеристика качества, то выбираем для фактора горизонтальную ось, а для характеристики качества – вертикальную. Чертится график, наносятся все необходимые данные, например, такие как название диаграммы, интервал времени, число пар данных,

названия и единицы измерения для каждой оси, имя человека, который делал эту диаграмму.

Пример 1. Фирма-производитель пластиковых емкостей, изготавливаемых методом литья под давлением, сталкивается с трудностями из-за дефектных емкостей, имеющих слишком тонкие стенки. Было высказано подозрение, что причина неподходящей толщины стенок заключается в вариации давления сжатого воздуха, которое каждый день меняется. В таблице 1 приведены данные о давлении воздуха и проценте дефектов. Построим диаграмму рассеяния по предложенным данным.

Таблица 1

Данные о давлении и проценте дефектов пластиковых емкостей

Дата	Давление, кгс/см ²	Процент дефектов	Дата	Давление, кгс/см ²	Процент дефектов
Окт. 1	8,6	0,889	Окт. 22	8,7	0,892
2	8,9	0,884	23	8,5	0,877
3	8,8	0,874	24	9,2	0,885
4	8,8	0,891	25	8,5	0,866
5	8,4	0,874	26	8,3	0,896
6	8,7	0,886	29	8,7	0,896
9	9,2	0,911	30	9,3	0,928
10	8,6	0,912	31	8,9	0,886
11	9,2	0,895	Ноя. 1	8,9	0,908
12	8,7	0,896	2	8,3	0,881
15	8,4	0,894	5	8,7	0,882
16	8,2	0,864	6	8,9	0,904
17	9,2	0,922	7	8,7	0,912
18	8,7	0,909	8	9,1	0,925
19	9,4	0,905	9	8,7	0,875

Влияющей переменной – фактором – является давление, обозначим его x , зависимая переменная – процент дефектов – y . Находим: $x_{\min}=8,2$; $x_{\max}=9,4$; $y_{\min}=0,864$; $y_{\max}=0,928$. Горизонтальную ось можно разметить от 8 до 9,5, вертикальную – от 0,86 до 0,94.

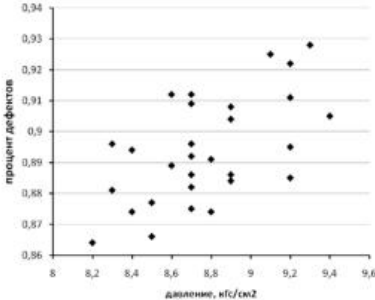
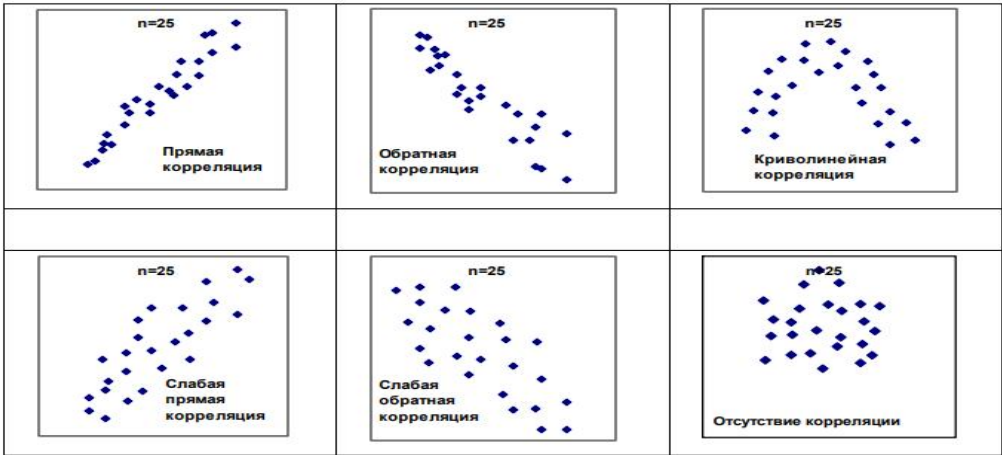


Рис. 1. Диаграмма рассеяния для давления сжатого воздуха и процента дефектов

Возможны многочисленные варианты скоплений точек, например:



Пример 1.

Изучается прочность на разрыв бумажной продукции. Предположили, что прочность зависит от количества твердой древесины в пульпе. На опытной установке изготовили 24 образцов, которые были испытаны на прочность на разрывной машине. Получили 24 пар данных, согласно этапам построения диаграммы разброса, их следует расположить в таблицу:

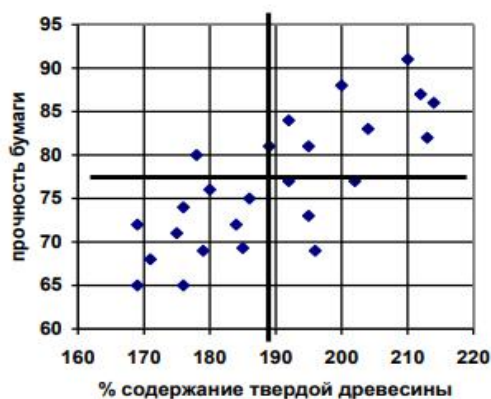
Таблица 1.

Распределение значений прочности

x	171	175	169	185	179	192	189	178	176	176	180	192
y	68	71	65	70	68	84	81	80	74	65	76	77
x	184	186	196	169	195	202	204	200	195	210	212	214
y	72	75	69	72	73	77	83	88	81	91	87	86

Далее определяем систему координат: на оси x отложим значения содержания твердой древесины в пульпе, а на оси y – полученные значения прочности бумаги. Для выбора начала координат нашли минимальные значения рассматриваемых параметров. Координатами начальной точки являются $x=160$, $y=60$. Наметили несколько градаций, для удобства максимальные значения рабочих зон выбрали для $x=220$, а для $y=95$. Нанесли на диаграмму данные, сделали все необходимые обозначения.

Диаграмма приняла вид



По виду полученной диаграммы можно предположить наличие слабой прямолинейной корреляции.

Задание для самостоятельной работы:

При точении валика из жаропрочной стали, проводится подогрев заготовки, с целью улучшения производительности процесса и качества обработанной поверхности. Предположили, что разность температурного поля оказывает существенное влияние на качественные характеристики обработанной поверхности. В лабораторных условиях проводится исследование влияния температуры разогрева заготовок, диапазон изменения температуры от 200 до 350 C⁰, с шагом в 5 C⁰. Данные о количестве несоответствующих деталей при различных температурах взять из таблицы случайных чисел, из столбца, указанного преподавателем.

2. Построим диаграмму рассеяния, иллюстрирующую взаимосвязь переменных x и y . Исходные данные.

Таблица 1. Исходные данные для построения диаграммы рассеяния.

x	y

5,8	5,4
7,2	6,3
9,1	12,7
14,8	16,8
16,2	17,1
13,7	19,2
13,2	16,3
17,6	23,2
13,2	19,3
14,2	16,3
20,5	31,2
21,6	35,1

Диаграмма рассеяния имеет вид:

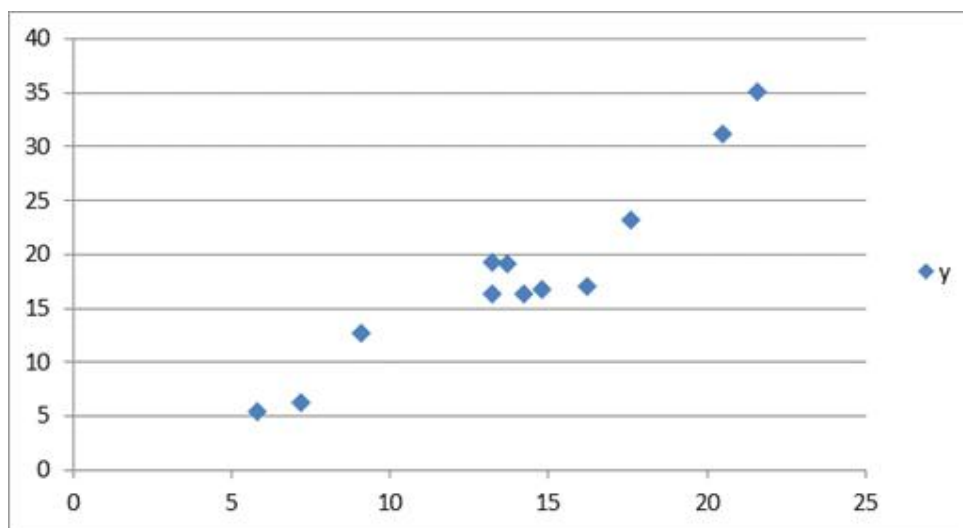


Рис. 1. Диаграмма рассеяния переменных x и y.

Задача №1.

Рассмотрим торговую точку. Проиллюстрируем зависимость количество проданной бутилированной воды в магазине и температурой воздуха в городе. В течение 30 дней проводились наблюдения за этими величинами и результаты заносились в таблицу:

X, °C	Y, л	X, °C	Y, л	X, °C	Y, л
19	26	26	42	29	61
18	24	25	44	27	58
21	29	26	48	26	50
23	32	27	52	24	41
20	25	27	54	25	45
22	33	24	46	23	38
24	32	25	50	24	35
25	31	27	55	26	43
25	34	28	50	28	52
25	38	28	55	29	56

Диаграмма разброса:

Коэффициент корреляции $r = 0.891973$, следовательно, существует прямая корреляция между случайными величинами.

Задача №2.

Построим диаграмму разброса случайных величин X и Y, представленных в таблице, если X- стоимость в рублях килограмма пшеничной муки, Y -стоимость в рублях килограмма хлебобулочных изделий из пшеничной муки. Определить вид корреляции между величинами X и Y.

Год	X, руб	Y, руб
1998	3,7	6,4
1999	8,0	11,0
2000	8,1	12,2
2001	8,5	13,7
2002	8,0	14,4
2003	11,4	18,7
2004	13,1	21,6
2005	12,0	22,2
2006	12,8	25,0
2007	17,4	30,7
2008	21,5	39,3
2009	19,5	39,7

Диаграмма разброса:

Коэффициент корреляции $r = 0.985710479$, следовательно, корреляция прямая сильная.

Задача №3.

Рассмотрим связь между числом учащихся X (в процентах) сделавших прививку от гриппа, и числом учащихся Y (в процентах) , заболевших гриппом в 40 школах города. Данные отражены в таблице:

$X, \%$	$Y, \%$	$X, \%$	$Y, \%$	$X, \%$	$Y, \%$	$X, \%$	$Y, \%$
20	35	23	31	17	40	24	29
21	35	39	12	21	34	33	17
15	42	19	37	31	21	17	41
18	38	31	22	27	28	28	28
36	16	25	23	30	26	37	15
35	17	20	33	25	23	19	35
40	12	38	14	36	18	34	20
16	40	35	19	23	30	29	27
18	36	26	28	32	20	20	34
24	31	32	19	22	30	38	15

Диаграмма разброса:

Коэффициент корреляции $r = -0,96115$, следовательно, существует обратная корреляция между случайными величинами.

Задача №4

На основе проведенных нами измерений роста и веса одноклассников построим диаграмму разброс. Определим вид корреляции между этими величинами

X , рост см	Y , вес кг	X , рост см	Y , вес кг
175	70	168	56
162	55	158	57
165	67	173	61
172	61	176	63
152	52	175	65
168	60	175	63
160	45	162	51
165	52	174	62
172	70	195	70
169	58	160	60

Диаграмма разброса:

Коэффициент корреляции $r = 0,71136$, следовательно, корреляционная зависимость между весом и ростом существует. Эта корреляция является прямой.

