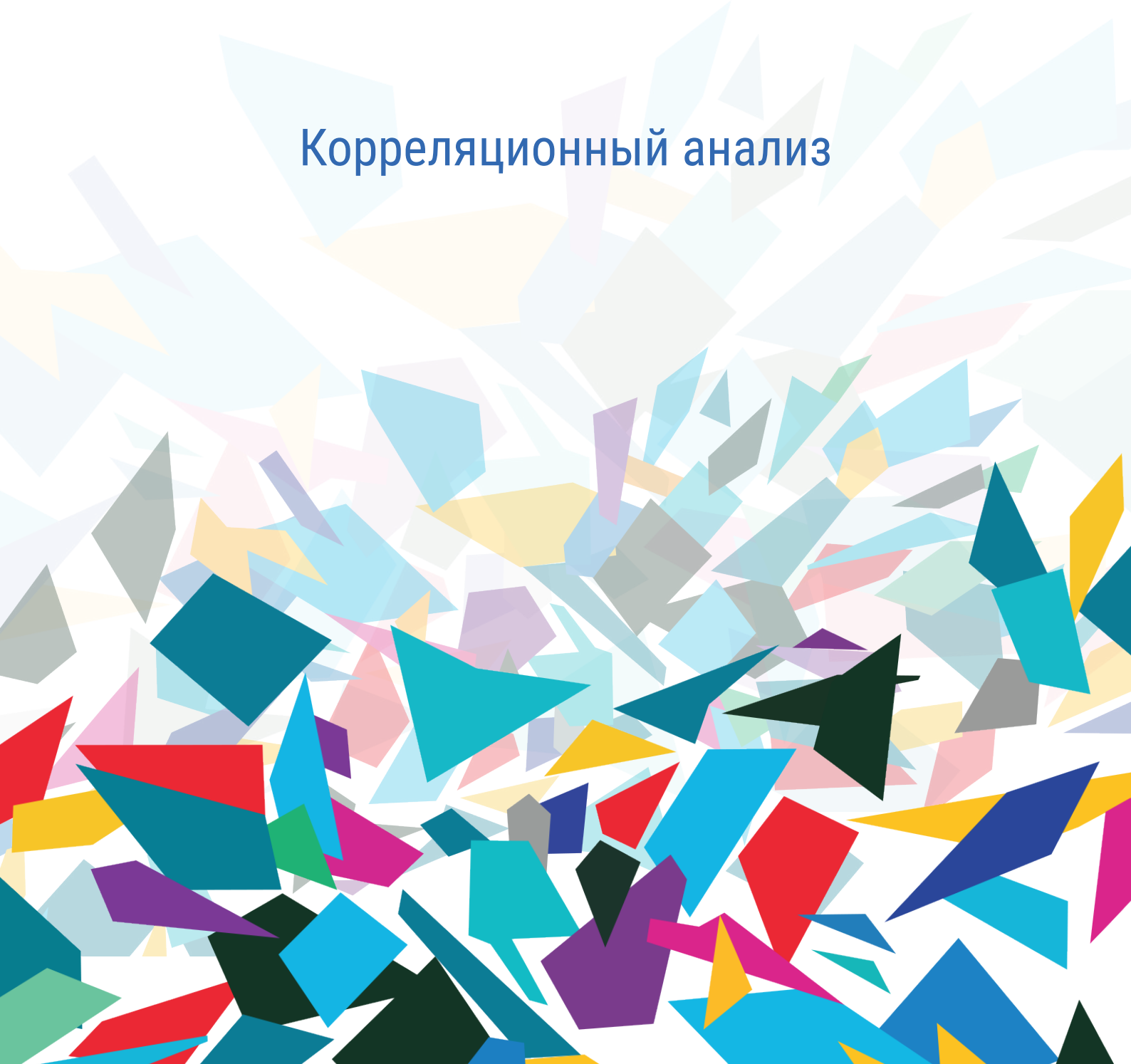
Лекция 17

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Корреляционный анализ



Здравствуйте!

Основными вопросами изучения являются: Понятие корреляции, видов корреляции, в частности, коэффициент КОРЕЛЯЦИИ, RПирсона, Спирмена и Кендала.

Статистика разработала множество методов изучения связей, выбор которых зависит от целей исследования и поставленных задач. Признаки по значению для изучения взаимосвязи делятся на два класса. Признаки, обуславливающие изменения других, связанных с ними признаков, называются факторными (X)-независимый фактор. Признаки, изменяющиеся под действием факторных признаков, называются результативными (Y)-зависимый фактор.

Существующие между явлениями формы и виды связей весьма разнообразны по своей классификации. Предметом статистики являются только такие из них, которые имеют количественный характер и изучаются с помощью количественных методов. Интерпретация результатов во многом зависит от того, насколько правильно были проведены расчёты, а также обоснованы применяемые методы. Одним из наиболее распространенных видов анализа статистической информации является корреляционный анализ. Его сущность заключается в поиске связи между двумя или более переменными. Причем наличие такой связи в первую очередь характеризуется тем, насколько сильно она выражена. Мерой тесноты двух коррелирующих величин является некоторый критерий, который получил название коэффициента корреляции. Этот коэффициент обозначается латинской букой r,и может принимать значения от +1 до -1. Чем ближе модуль коэффициента корреляции к единице, тем более сильной является связь между измеряемыми величинами. Отсутствие связи характеризуется коэффициентом корреляции равным 0 или близким к нему значением. Существует следующая градация силы связи, представленная шкалой Чертока.

Таблица 17.1 Шкала Чертока

Значение	Характеристика
$0 < \|r\| < 0,1$	Связь практически отсутствует
$0,1 < \|r\| < 0,3$	Слабая связь
$0,3 < \|r\| < 0,5$	Умеренная связь
$0,5 < \|r\| < 0,7$	Связь средней силы
$0,7 < \|r\| < 0,9$	Сильная связь
$0,9 < \|r\| < 1$	Очень сильная связь

Есть мнение, что в социологических исследованиях значения коэффициентов корреляции выше 0,5 встречаются не очень часто, поэтому можно принимать во внимание те из них, которые равны или превышают 0,3, т. е. характеризуют умеренную взаимосвязь признаков. Если коэффициент корреляции отрицательный, это означает наличие противоположной связи: чем выше значение одной переменной, тем ниже значение другой. Широкое применение корреляционный анализ нашел в гуманитарных науках, в том числе и при изучении социологических проблем. Объясняется это в первую очередь тем, что исследуемые социологической наукой вопросы могут включать в себя большое число влияющих факторов. То есть определенному значению одной переменной, соответствует целый комплекс значений другой, представляющий собой

ряд распределения, причем при изменении данной величины меняется ряд распределения и его среднее. Если каждую пару значений этих величин изобразить на плоскости в декартовой системе координат с помощью точек, то наносимые точки расположатся в виде «облака», называемое корреляционным полем.

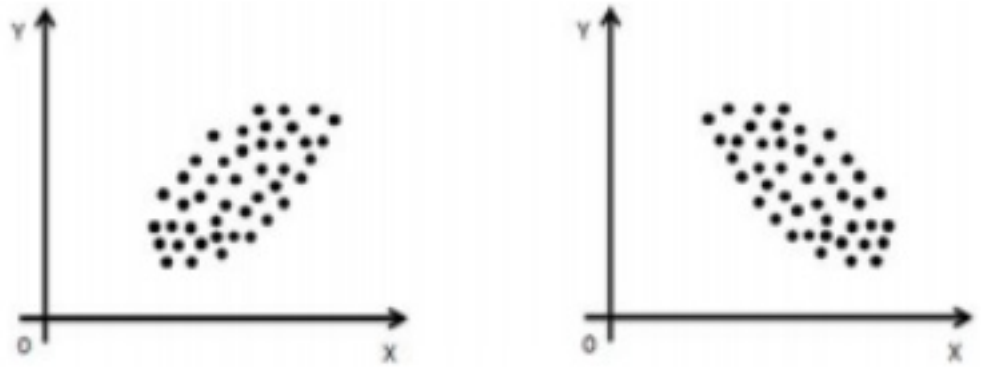


Рисунок 17.1 Корреляционное поле (слева положительная корреляция, справа отрицательная)

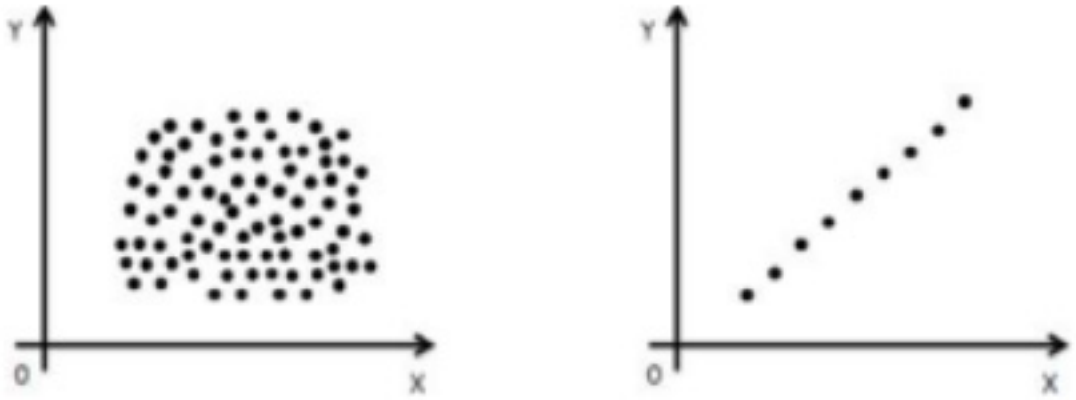


Рисунок 17.2 Корреляционное поле (слева –отсутствие связи, справа – функциональная связь).

Корреляционная зависимость не является абсолютно полной и точной. Она включает в себя множественность причин и следствий разных порядков. А любая социологическая проблема представляет собой явление, которое зависит от большого числа причин, действующих с разной силой. Изучая влияние одной переменной на другую, мы выделяем один фактор, но на зависимую величину оказывают влияние и многие другие, чем и обусловлен характер корреляционной зависимости. Другой причиной удобства использования корреляционного анализа в социологии является применимость его для переменных, относящихся к различным шкалам.
Различают параметрические (Пирсона) и непараметрические (Спирмена, Кендалла, тау) способы подсчёта коэффициента корреляции.
Для обозначения параметрического коэффициента корреляции Пирсона обычно используется обозначение r, для рангового коэффициента корреляции Спирмена – обозначение ρ.

Более детально методы вычисления коэффициента корреляции зависит от вида шкалы, которой относятся переменные, представленные в таблице 17.2.

Таблица 17.2. Методы вычисления коэффициента корреляции

Типы шкал		Мера связи
Переменная X	Переменная Y	
Интервальная (или отношений)	Интервальная (или отношений)	Коэффициент Пирсона
Ранговая, интервальная (или отношений)	Ранговая, интервальная (или отношений)	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла
Дихотомическая	Дихотомическая	Коэффициент φ (фи), четырёх полевая корреляция
Дихотомическая	Ранговая	Рангово-бисериальный коэффициент

Линейный коэффициент корреляции r-Пирсона

Коэффициент корреляции Пирсона (r-Пирсона) применяется для исследования взаимосвязи двух переменных, измеренных в метрических шкалах на одной и той же выборке. Он позволяет определить, насколько пропорциональная изменчивость двух переменных.
Данный коэффициент разработали Карл Пирсон, Фрэнсис Эджуорт и Рафаэль Уэлдон в 90-х годах XIX века. Коэффициент корреляции изменяется в пределах от минус единицы до плюс единицы.

Коэффициент корреляции r-Пирсона характеризует существование линейной связи между двумя величинами. Если связь криволинейная, то он не будет работать.

Чтобы приступать к расчетам коэффициента корреляции r-Пирсона необходимо выполнение следующих условий:
Исследуемые переменные X и Y должны быть распределены нормально.
Исследуемые переменные X и Y должны быть измерены в интервальной шкале или шкале отношений.
Количество значений в исследуемых переменных X и Y должно быть одинаковым.
При расчете коэффициент линейной корреляции Пирсона используется специальная формула.
Величина коэффициента корреляции варьируется от 0 до 1.

Слабыми сторонами линейного коэффициента корреляции Пирсона являются:
Неустойчивость к выбросам.
С помощью коэффициента корреляции Пирсона можно определить только силу линейной взаимосвязи между переменными, другие виды взаимосвязей выявляются методами регрессионного анализа.
Пример расчета коэффициента корреляции Пирсона.
Рассмотрим пример использования коэффициента корреляции Пирсона.
Например, нам необходимо определить взаимосвязь двух переменных агрессивности и IQ у студентов по полученным данным тестирования.
Данные сведем в одну таблицу:

Таблица 17.3а

№	Данные по агрессивности ($X_{агр}$)	Данные по IQ (Y_{IQ})
1	24	100
2	27	115
3	26	117
4	21	119
5	20	134
6	31	94
7	26	105
8	22	103
9	20	111
10	18	124
11	30	122
12	29	109
13	24	110
14	26	86

1. Вычисляем сумму значений X_{aqr} агрессивности и Y_{IQ}

$X_{aqr} = 344$

$Y_{IQ} = 1549$

2. Вычисляем среднее арифметическое для X_{aqr} и Y_{IQ}

$X_{aqr} = 24,6$

$Y_{IQ} = 110,5$

3. Вычисляем для каждого испытуемого отклонения от среднего арифметического для X_{aqr} и $Y_{IQ.}$

Таблица 17.3б

№	$\overline{X}_{aqr} - X_{aqr}$	$\overline{Y}_{IQ} - Y_{IQ}$
1	0,6	10,6
2	-2,4	-4,4
3	-1,4	-6,4
4	3,6	-8,4
5	4,6	-23,4
6	-6,4	16,6
7	-1,4	5,6
8	2,6	7,6
9	4,6	-0,4
10	6,6	-13,4
11	-5,4	-11,4
12	-4,4	1,6
13	0,6	0,6
14	-1,4	24,6

4. Затем мы возводим в квадрат каждое отклонение:

Таблица 17.3в

№	$(\overline{X}_{aqr} - X_{aqr})^2$	$(\overline{Y}_{IQ} - Y_{IQ})^2$
1	0,36	112,36
2	5,76	19,36
3	1,96	40,96
4	12,96	70,56
5	21,16	547,56
6	40,96	275,56
7	1,96	31,36
8	6,76	57,79
9	21,16	0,16
10	43,56	179,56
11	29,16	129,96
12	19,36	2,56
13	0,36	0,36
14	1,96	605,16

5. Потом рассчитываем сумма квадратов отклонений: $\sum(\overline{X}_{aqr} - X_{aqr})^2$ и $\sum(\overline{Y}_{IQ} - Y_{IQ})^2$

$$\sum(\overline{X}_{aqr} - X_{aqr})^2 = 207,44$$

$$\sum(\overline{Y}_{IQ} - Y_{IQ})^2 = 2073,24$$

6. Рассчитываем для каждого наблюдения произведение разности среднего арифметического и значения.

Таблица 17.3г

№	$(\overline{X}_{aqr} - X_{aqr}) * (\overline{Y}_{IQ} - Y_{IQ})$
1	6,36
2	10,56
3	8,96
4	-30,24
5	-107,64
6	-106,24
7	-7,84
8	19,76
9	-1,84
10	-88,44
11	61,56
12	-7,04
13	0,36
14	-34,44

7. Рассчитываем сумму $(\overline{X}_{aqr} - X_{aqr}) * (\overline{Y}_{IQ} - Y_{IQ})$.

$\sum(\overline{X}_{aqr} - X_{aqr}) * (\overline{Y}_{IQ} - Y_{IQ}) = -276,16$

8. Подставляем полученные значения σX_{aqr}, σY_{IQ}, $\sum (\overline{X}_{aqr} - X_{aqr}) * (\overline{Y}_{IQ} - Y_{IQ})$ в формулу коэффициента корреляции Пирсона.

В соответствии с таблицей значений величин коэффициента корреляции делаем вывод о том, что rXaqrYIQ= -0.421 это слабая по силе отрицательная корреляция.

Коэффициент ранговой корреляции r-Спирмена

Коэффициент ранговой корреляции r-Спирмена применяется для исследования корреляционной взаимосвязи между двумя ранговыми переменными. Коэффициент ранговой корреляции r-Спирмена может быть вычислен двумя способами:

1. Применением формулы коэффициента корреляции Пирсона. Важно! Переменные предварительно должны быть ранжированы.
2.Использование упрощенной формулы коэффициента корреляции:

Формула 17.1 $$r_S = 1 - \frac{6\Sigma_i d_i^2}{N\,(N^2 - 1)}$$

Важно! Формула используется при отсутствии повторяющихся рангов.
Пример расчета коэффициента корреляции r-Спирмена.
Рассмотрим расчет коэффициента корреляции r-Спирмена на примере. Допустим у нас есть данные на 14 студентов одной группы по уровню интеллекта (IQ) и время решения серии логических заданий (X).

Таблица 17.4а

№	Уровень интеллекта (IQ)	Время решения логических задач в секундах (X)
1	100	154
2	118	123
3	112	120
4	97	213
5	99	200
6	103	187
7	102	155
8	132	100
9	122	114
10	121	115
11	115	107
12	117	176
13	109	143
14	111	111

1. Проранжируем полученные данные по столбцу (переменной) IQ и по столбцу (переменной) X.

Таблица 17.4б

№	ранг IQ	ранг X
1	3	9
2	11	7
3	8	6
4	1	14
5	2	13
6	5	12
7	4	10
8	14	1
9	13	4
10	12	5
11	9	2
12	10	11
13	6	8
14	7	3

2. Вычислим разность рангов по каждому случаю.
Таблица 17.4в

№	delta = ранг IQ — ранг X
1	-6
2	4
3	2
4	-13
5	-11
6	-7
7	-6
8	13
9	9
10	7
11	7
12	-1
13	-2
14	4

3. Возведем полученную на втором шаге разность в квадрат.
Таблица 17.4г

№	$delta^2$
1	36
2	16
3	4
4	169
5	121
6	49
7	36
8	169
9	81
10	49
11	49
12	1
13	4
14	16

4. Найдем сумму квадратов разностей:
36+16+4+169+121+49+36+169+81+49+49+1+4+16 = 800
5. Подставим имеющиеся значения в формулу коэффициента корреляции r-Спирмена.

$$r_S = 1 - \frac{6 \cdot 800}{14 \cdot (196 - 1)} = -0,76$$

между уровнем IQ и агрессивностью есть сильная отрицательная связь со значением -0,76.

Коэффициент ранговой корреляции Кендалла.
Коэффициент ранговой корреляции T -Кендалла является альтернативой методу определения корреляции r-Спирмана. Он предназначен для определения взаимосвязи между двумя ранговыми переменными.
Интерпретация результатов вычисления коэффициент ранговой корреляции T -Кендалла определяется как разность вероятностей совпадения и инверсии в рангах.

Для одних и тех же значений переменных значения коэффициента корреляции r-Спирмена будет всегда немного больше, чем значения коэффициента ранговой корреляции T -Кендалла, тогда как уровень значимости будет одинаков или же у коэффициента корреляции T -Кендалла будет

немного больше.

Формула вычисления коэффициента ранговой корреляции Т -Кендалла отличается от формулы коэффициента корреляции r-Пирсона, и может быть выражена как:

Формула 17.2 $$\tau = \frac{P\,(p)-P(q)}{N\frac{(N-1)}{2}}$$

где P(p) — число совпадений, P(q) — число инверсий, N — объем выборки
В упрощенном виде формулу коэффициента корреляции Кендалла можно записать как:
Формула 17.3

$$\tau = \frac{4P}{N(N-1)} - 1$$

При наличии связанных рангов формула изменяется с учетом поправки на связанные ранги:
Формула 17.4

$$\tau = \frac{P\,(p)-P(q)}{\sqrt{\left[N\frac{(N-1)}{2}\right]-K_x}\sqrt{\left[N\frac{(N-1)}{2}\right]-K_y}}$$

где P(p) — число совпадений, P(q) — число инверсий, N — объем выборки, Kx— поправка на связи рангов переменной X, Ky — поправка на связи рангов переменной Y.
Kx = 0.5∑i ∫i (∫i -1), где i — количество групп связей по X ,∫i — численность группы X
Ky = 0.5∑i ∫i (∫i -1), где i — количество групп связей по Y, ∫i — численность группы Y.

Коэффициенты частной корреляции позволяют изучать связи между признаками при элиминировании, т.е. и влияния некоторых других признаков. Если устраняется влияние одного признака, то говорят о частных Коэффициентах корреляции первого порядка. Они выражаются через обычные коэффициенты парной корреляции. Логика частной корреляции такова: если при устранении некоторого признака коэффициент корреляции двух данных признаков увеличивается, то такой признак ослабляет связь, если же коэффициент корреляции уменьшается, то устраняемый признак в определенной мере обуславливает связь. (В предельном случае, если устранение признака обращает коэффициент корреляции в нуль, то данный признак обуславливает связь данных признаков, т.е. это связь сопутствия).

При изучении корреляции между производительностью труда и возрастом рабочих была установлена положительная связь. На производительность влияет и стаж работы, который оказывается в положительной корреляции и с возрастом, и с производительностью. При элиминировании стажа оказалось, что связь между производительностью и возрастом отрицательная, а между производительностью труда и стажем (при элиминировании возраста) – положительная и еще более тесная. Если устраняется влияние двух признаков, то говорят о частных Коэффициентах корреляции второго порядка. Они в свою очередь выражаются через коэффициенты частной корреляции первого порядка и т.д.
Корреляционная таблица (таблица сопряженности признаков) – один из основных способов описания корреляционных связей между признаками, используемых для упорядочения информации о распределении изучаемой совокупности индивидов по двум признакам. Корреляционная таблица имеет прямоугольную форму, число строк ее n – определяется количеством значений одного признака, а число столбцов m – количеством значений другого.

Общий вид таблицы сопряженности

X	Y						Маргиналы по строкам
	1	2	...	j	...	c	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2.}$
...	...	...	...	...	...	...	...
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	$n_{i.}$
...	...	...	...	...	...	...	...
r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	$n_{r.}$
Маргиналы по столбцам	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.c}$	n

На пересечении, например, второй строки и третьего столбца в таблице проставляется число индивидов, у которых первый признак принимает второе значение из своего списка, а второй признак – третье из своего. Таблица имеет n x m внутренних клеток. Кроме того, выделяются два маргинала (на полях – правом и нижнем). Первый маргинал это m + 1-ый столбец, заполненный числами индивидов, у которых первый признак принимает свое первое значение (независимо от того, какое значение принимает второй признак, эта сумма элементов первой внутренней строки), второе значение и т.д. до n-ого. Второй маргинал это n + 1-ая строка, заполненная суммами элементов соответствующих столбцов. Сумма элементов каждого маргинала равна числу индивидов. Такого рода распределения называют двумерными. Если n=1,
то говорят об одномерном распределении (оно показывает, как распределены индивиды по одному, в данном случае второму признаку). Изучают и трехмерные распределения: для каждого значения третьего признака составляют свои двумерные распределения – по первому и второму признакам и т.д. Таким образом, основной формой представления является двумерная Корреляционная таблица. Характер распределения индивидов по ее клеткам определяется характером связи между признаками. Поэтому по эмпирической таблице восстанавливают характер связи. Если связи нет, то число индивидов, попадающих в данную клетку таблицы, равно произведению маргиналов строки и столбца с соответствующими номерами, деленному на число всех индивидов. Таблицу, заполненную такими частотами, называют теоретической. Если связь есть, то эмпирическая таблица отличается от теоретической. Мерой отличия, характеризующей связь, является критерий Пирсона хи-квадрат о котором мы говорили ранее. Анализ таблиц сопряженности – то есть распределений по двум и более переменным, начинается с установления наличия связи между переменными.
Анализ основан на значениях клеточных частот. Если таблица сопряженности квадратная, т.е. число строк равно числу столбцов и все частоты в диагональных клетках не равны нулю, а остальные – имеют нулевое значение, то между переменными имеется полная связь и наоборот, если данные распределены достаточно равномерно по клеткам таблицы, то имеется слабая связь, либо связь отсутствует. При любом числе строк и столбцов факт наличия или отсутствия связи устанавливается с помощью критерия хи-квадрат.

Как дополнение можно также рассмотреть следующее понятие, напоминаю, что информация несет информативный характер.
Корреляция рядов динамики – метод изучения связи между показателями, представленными их

значениями в последовательные моменты или периоды времени. Коэффициент корреляции между непосредственно заданными значениями показателей может дать превратное представление об их связи, поскольку он может отразить всего лишь совпадение их общей тенденции изменения. В этом случае корреляция называется ложной. Это породило правило: определить корреляцию не между самими значениями показателя, а между их отклонениями от линий, выражающих для каждого ряда тенденцию.
Классификация тесноты связи по значению модуля коэффициента линейной корреляции, аналогична вышесказанного.

Коэффициент корреляции в применении к рядам динамики связан с параллельностью вариации явлений: если общий характер вариации двух рядов (т.е. гладкая и периодическая составляющая тренда) является одинаковым, то показатель тесноты связи будет большим. Ясно, что одинаковые вариации могут встречаться и в силу чисто случайных обстоятельств, поэтому необходим упомянутый всесторонний логико-теоретический анализ.

Корреляционная зависимость между уровнями взаимосвязанных рядов динамики.
При изучении развития явления во времени часто возникает необходимость оценить степень взаимосвязи в изменениях уровней 2-х или более рядов динамики различного содержания, но связанных между собой.
Коррелирование уровней динамических рядов с применением парного коэффициента корреляции правильно показывает тесноту связи лишь в том случае, если в каждом из них отсутствует автокорреляция. Наличие зависимости между последующими и предшествующими уровнями динамического ряда в статистической литературе называют автокорреляцией.

Поэтому прежде, чем коррелировать ряды динамики по уровням, необходимо проверить каждый из рядов на наличие или отсутствие в них автокорреляции.
Применение методов классической теории корреляции в динамических рядах связано с некоторыми особенностями. Прежде всего, это наличие для большинства динамических рядов зависимости последующих уровней от предыдущих.

Коэффициент автокорреляции вычисляется по непосредственным данным рядов динамики, когда фактические уровни одного ряда рассматриваются как значения факторного признака, а уровни этого же ряда со сдвигом на один период, принимаются в качестве результативного признака (этот сдвиг называется лагом). Коэффициент автокорреляции рассчитывается на основе формулы коэффициента корреляции для парной зависимости.

Но существуют и ограничения использования корреляционного анализа. Применение возможно при наличии достаточного количества наблюдений для изучения. На практике считается, что число наблюдений должно не менее чем в 5-6 раз превышать число факторов (также встречается рекомендация использовать пропорцию, не менее чем в 10 раз превышающую количество факторов). В случае если число наблюдений превышает количество факторов в десятки раз, в действие вступает закон больших чисел. Необходимо, чтобы совокупность значений всех факторных и результативного признаков подчинялась многомерному нормальному распределению. В случае если объём совокупности недостаточен для проведения формального тестирования на нормальность распределения, то закон распределения определяется визуально на основе корреляционного поля. Если в расположении точек на этом поле наблюдается линейная тенденция, то можно предположить, что совокупность исходных данных подчиняется нормальному закону распределения. Исходная совокупность значений должна быть качественно однородной. Сам по себе факт корреляционной зависимости не даёт основания утверждать, что

одна из переменных предшествует или является причиной изменений, или то, что переменные вообще причинно связаны между собой, а не наблюдается действие третьего фактора. Область применения. Данный метод обработки статистических данных весьма популярен в экономике и социальных науках (в частности в психологии и социологии), хотя сфера применения коэффициентов корреляции обширна: контроль качества промышленной продукции, металловедение, агрохимия, гидробиология, биометрия и прочие. В различных прикладных отраслях приняты разные границы интервалов для оценки тесноты и значимости связи. Популярность метода обусловлена двумя моментами: коэффициенты корреляции относительно просты в подсчете, их применение не требует специальной математической подготовки. В сочетании с простотой интерпретации, простота применения коэффициента привела к его широкому распространению в сфере анализа статистических данных.
Подведем итоги изученного материала, нам теперь известно, что существует шкала связи коэффициента корреляции, а также что существует прямая и обратная связь, которую можно изобразить графически в виде облаков. Что метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные: если обе переменные измерены в интервальной и количественной шкалах необходимо использовать коэффициент корреляции Пирсона, если одна из двух переменных имеет порядковую шкалу, либо не является нормально распределённой, необходимо использовать ранговую корреляцию Спирмена или коэффициент корреляции (тау) Кендалла. Кроме того, задача научного исследования состоит в отыскании причинных зависимостей, выявлении истинных причин. Но корреляция как формальное статистическое понятие сама по себе не вскрывает причинного характера связи. С помощью корреляционного анализа нельзя указать, какую переменную принимать в качестве причины, а какую – в качестве следствия, это дело исследователя. Иногда при наличии корреляционной связи ни одна из переменных не может рассматриваться причиной другой. В некоторых случаях возможна ложная корреляция (нонсенс-корреляция), т.е. чисто формальная связь между переменными, не находящая никакого объяснения и основанная лишь на количественном соотношении между ними.

Вывод:

Существует шкала связи коэффициента корреляции.
Существует прямая и обратная связь, которую можно изобразить графически.
Метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные: если обе переменные измерены в интервальной и количественной шкалах необходимо использовать коэффициент корреляции r-Пирсона, если одна из двух переменных имеет порядковую шкалу, либо не является нормально распределённой, необходимо использовать ранговую корреляцию Спирмена или коэффициент корреляции (тау) Кендалла.