Ковариация, дисперсия и корреляция

Выборочная и теоретическая ковариации

- Ковариация является мерой взаимосвязи между двумя переменными
- Если x и y - случайные величины, то **теоретическая ковариация** определяется как математическое ожидание произведения отклонений этих величин от их средних значений:

$$\mathbf{cov}(x,y) = \sigma_{xy} = E[(x-\mu_x)(y-\mu_y)]$$

•где μ_x и μ_y - теоретические средние значения x и y соответственно.

- При наличии n наблюдений двух переменных (x и y) **выборочная ковариация** между x и y задается формулой:

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Можно сказать, что ковариация **характеризует сопряженность вариации двух признаков и представляет собой статистическую меру взаимодействия двух случайных переменных**

- Если теоретическая ковариация неизвестна, то для ее оценки может быть использована **выборочная ковариация**, вычисленная по ряду наблюдений.

- Эта оценка будет иметь отрицательное смещение.
- Причина заключается в том, что выборочные отклонения измеряются по отношению к выборочным средним значениям величин x и y и **имеют тенденцию к занижению отклонений** от истинных средних значений.

- Можно рассчитать несмещенную оценку путем умножения выборочной оценки на **n / (n - 1)** .
- **Если x и y независимы, то их теоретическая ковариация равна нулю**.

Пример расчета ковариации

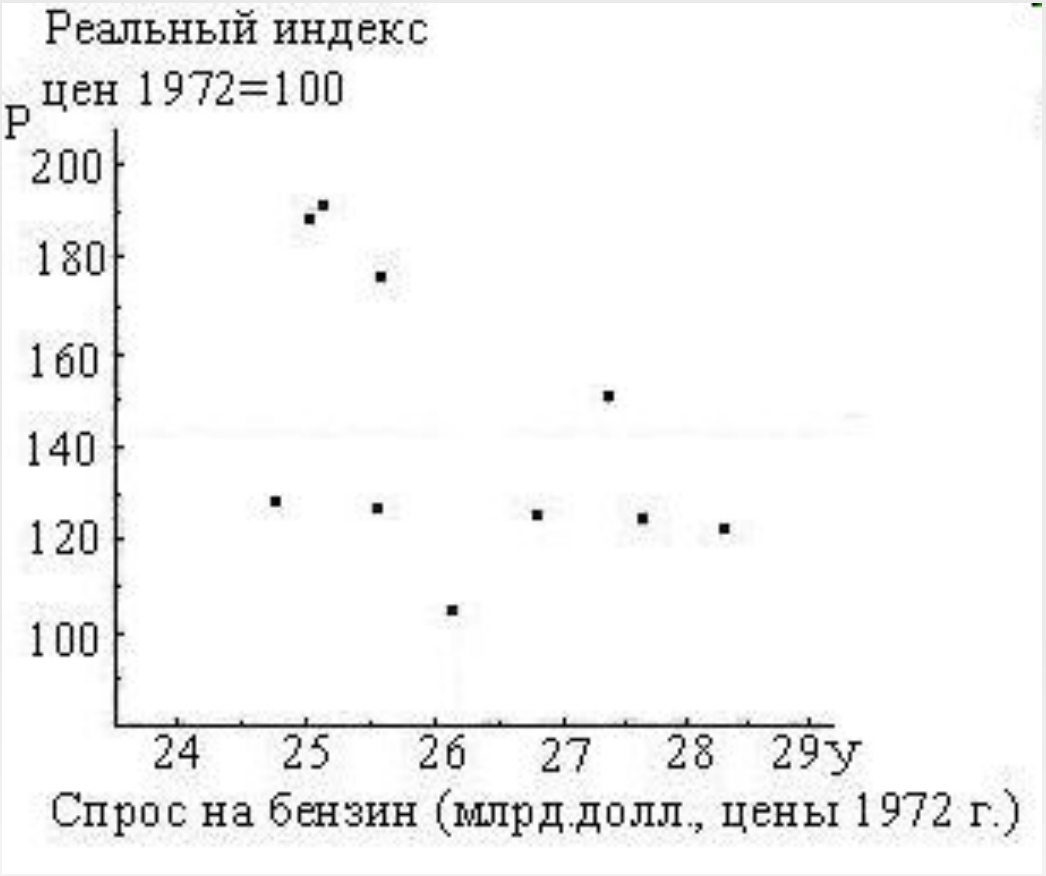
- Со времен нефтяного кризиса 1973 г. реальная цена на бензин, т.е. цена бензина, отнесенная к уровню общей инфляции, значительно возросла, и это оказало заметное воздействие на потребительский спрос.
- В период между 1963 и 1972 гг. потребительский спрос на бензин устойчиво повышался.
- Эта тенденция прекратилась в 1973 г., а затем последовали нерегулярные колебания спроса с незначительным его падением в целом.

Потребительские расходы на бензин и его реальная цена в США		
Год	Расходы (млрд. долл., цены 1972 г.)	Индекс реальных цен (1972 = 100)
1973	26,2	103,5
1974	24,8	127,0
1975	25,6	126,0
1976	26,8	124,8
1977	27,7	124,7
1978	28,3	121,6
1979	27,4	149,7
1980	25,1	188,8
1981	25,2	193,6
1982	25,6	173,9

- В таблице приведены данные о потребительском спросе и реальных ценах после нефтяного кризиса.

- Реальная цена вычислялась путем деления индекса номинальной цены на бензин, на общий индекс потребительских цен и умножения результата на 100.
- Индексы основаны на данных 1972 г.; индекс реальной цены показывает повышение цены бензина относительно общей инфляции начиная с 1972г.

Эти данные показаны в виде диаграммы рассеяния.



Можно видеть отрицательную связь между потребительским спросом на бензин и его реальной ценой.

- Показатель выборочной ковариации позволяет выразить данную связь единым числом.
- Для его вычисления мы сначала находим средние значения цены и спроса на бензин.

Наблюдение	p	y
1973	103,5	26,2
1974	127,0	24,8
1975	126,0	25,6
1976	124,8	26,8
1977	124,7	27,7
1978	121,6	28,3
1979	149,7	27,4
1980	188,8	25,1
1981	193,6	25,2
1982	173,9	25,6
Сумма	**1433,6**	**262,7**
Среднее	**143,36**	**26,27**

- Обозначив цену через **p** и спрос через **y**, определяем средние значения, которые оказываются равными соответственно **143,36 и 26,27.**

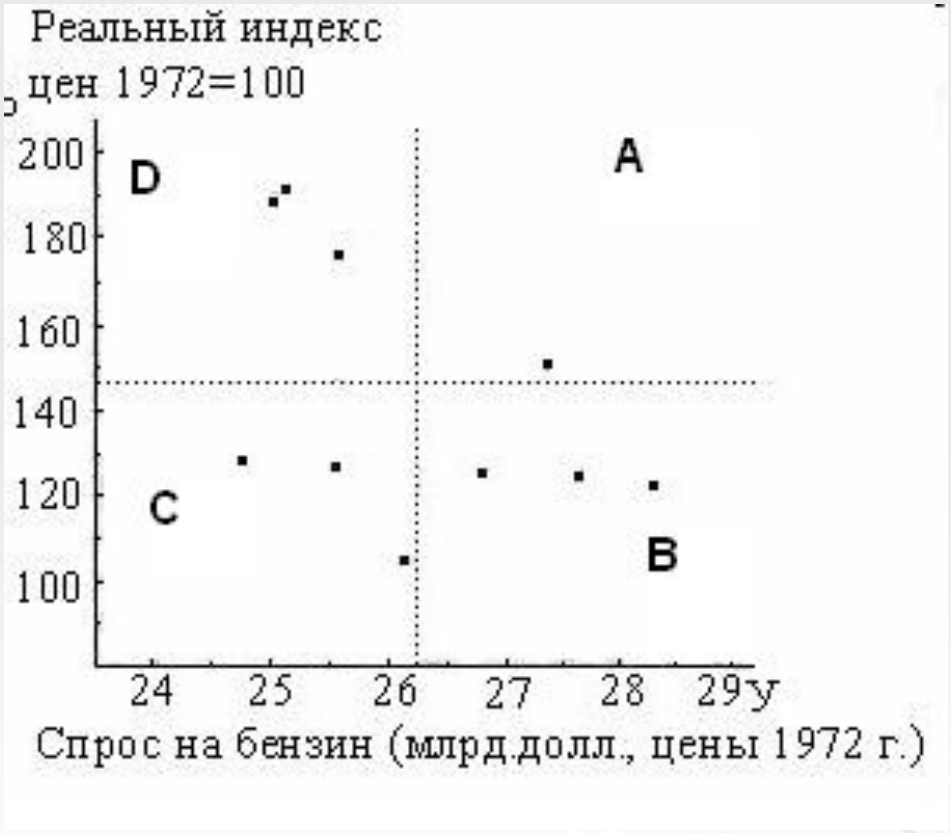
- Затем для каждого года вычисляем отклонение величин **p** и **y** от средних и перемножаем их.

Наблюдение	p	y	$(p-\bar{p})$	$(y-\bar{y})$	$(p-\bar{p})(y-\bar{y})$
1973	103,5	26,2	-39,86	-0,07	2,79
1974	127,0	24,8	-16,36	-1,47	24,05
1975	126,0	25,6	-17,36	-0,67	11,63
1976	124,8	26,8	-18,56	0,53	-9,84
1977	124,7	27,7	-18,66	1,43	-26,68
1978	121,6	28,3	-21,76	2,03	-44,17
1979	149,7	27,4	6,34	1,13	7,16
1980	188,8	25,1	45,44	-1,17	-53,16
1981	193,6	25,2	50,24	-1,07	-53,76
1982	173,9	25,6	30,54	-0,67	-20,46
Сумма	**1433,6**	**262,7**			**-162,44**
Среднее	**143,36**	**26,27**			**-16,24**

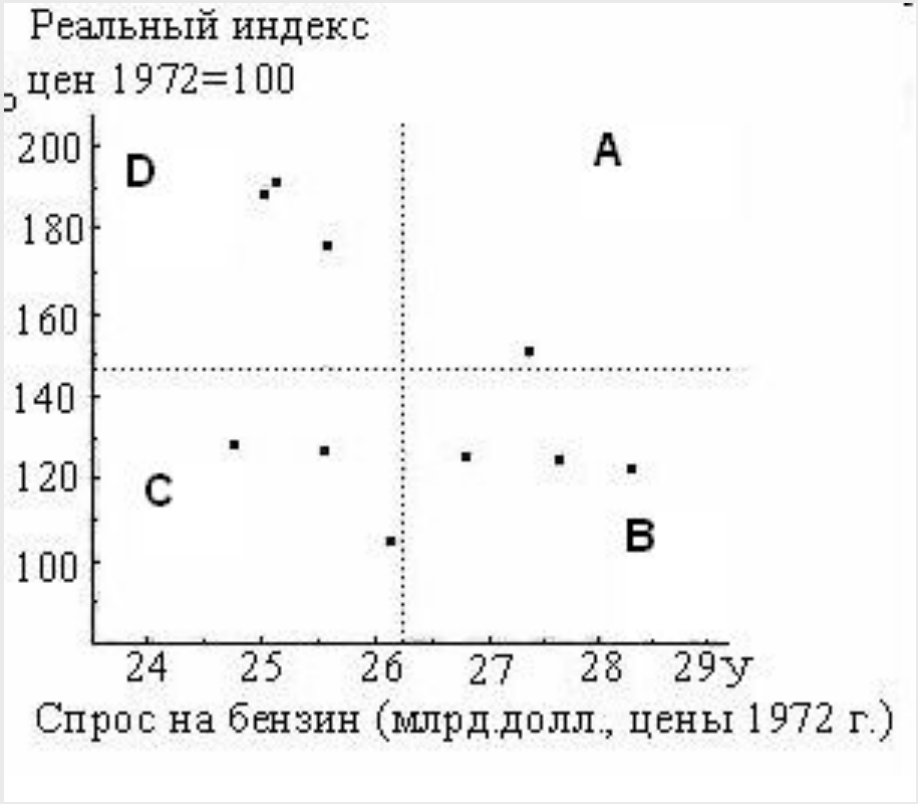
В нижней клетке последнего столбца определяется средняя величина (-16,24), она является значением выборочной ковариации.

- Ковариация в данном случае отрицательна.
- Так это и должно быть.
- Отрицательная связь, как это имеет место в данном примере, выражается отрицательной ковариацией, а положительная связь - положительной ковариацией.

- На рисунке диаграмма рассеяния наблюдений делится на четыре части вертикальной и горизонтальной линиями, проведенными через средние значения p и y соответственно.



- Пересечение этих линий образует точку , которая показывает **среднюю цену и средний спрос** за период, соответствующий выборке.



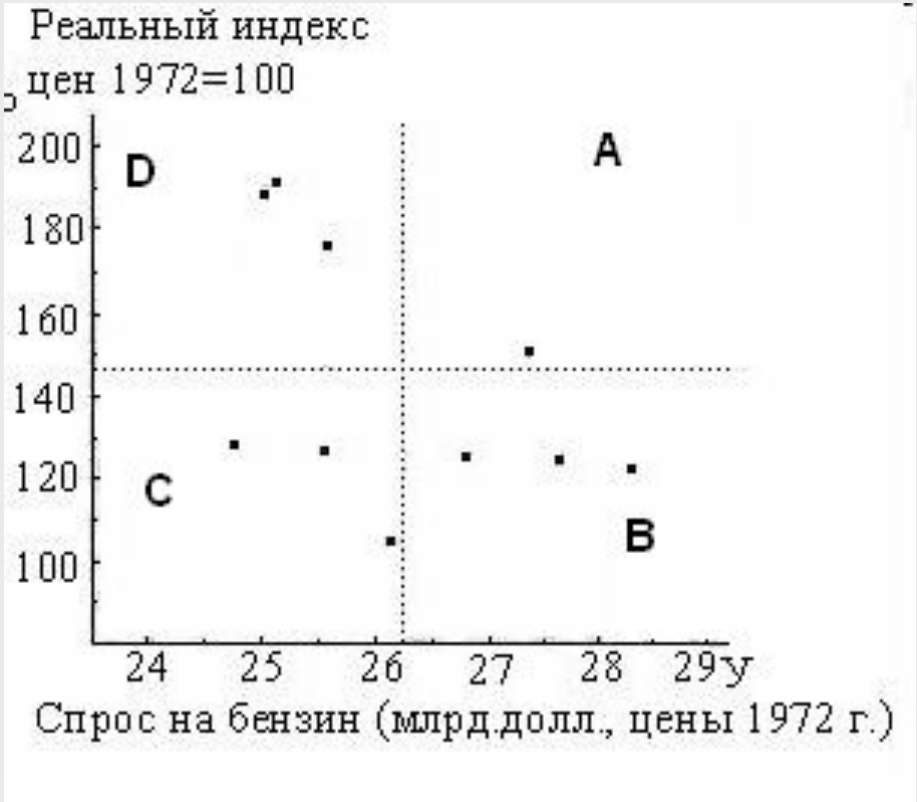
Для любого наблюдения, лежащего в квадранте **A**, значения реальной цены и спроса **выше** соответствующих средних значений.

Здесь $(p-\bar{p})$ и $(y-\bar{y})$ являются положительными, а поэтому должно быть положительным и

$$(p-\bar{p})(y-\bar{y})$$

Наблюдения дают положительный вклад в ковариацию.

В квадранте B наблюдения имеют реальную цену **ниже** средней и спрос **выше** среднего. Наблюдения дают отрицательный вклад в ковариацию.



В квадранте C как реальная цена, так и спрос **ниже** своих средних значений. Наблюдения дают положительный вклад в ковариацию.

В квадранте D реальная цена **выше** средней, а спрос **ниже** среднего. Наблюдения дают отрицательный вклад в ковариацию

- Поскольку выборочная ковариация является средней величиной произведения для 10 наблюдений, она будет **положительной, если положительные вклады будут доминировать над отрицательными, и отрицательной, если будут доминировать отрицательные вклады**.
- Положительные вклады исходят из квадрантов A и C, и ковариация будет, скорее всего, положительной, если основной разброс пойдет по наклонной вверх.

- Точно так же отрицательные вклады исходят из квадрантов B и D.
- Поэтому если основное рассеяние идет по наклонной вниз, как в данном примере, то ковариация будет, скорее всего, отрицательной.

Правила расчета ковариации

- Существует несколько правил, которые вытекают непосредственно из определения ковариации.
- Правило 1:

Если y = v + w, то

Cov(x, y) = Cov(x, v) + Cov(x, w).

- Допустим, имеются данные по 6 семьям: общий годовой доход (x); расходы на питание и одежду (y), расходы на питание (v), расходы на одежду (w). Естественно, y = v + w

Семья	Доход семьи (x)	Расходы на питание и одежду (y)	расходы на питание (v)	Расходы на одежду (w)
1	3000	1100	850	250
2	2500	850	700	150
3	4000	1200	950	250
4	6000	1600	1150	450
5	3300	1000	800	200
6	4500	1300	950	350

Семья	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(v-\bar{v})$	$(x-\bar{x})(v-\bar{v})$	$(w-\bar{w})$	$(x-\bar{x})(w-\bar{w})$
1	-883	-75	66250	-50	44167	-25	22083
2	-1383	-325	449583	-200	276667	-125	172917
3	117	25	2917	50	5833	-25	-2917
4	2117	425	899586	250	529167	175	370416
5	-583	-175	102083	-100	58333	-75	43750
6	617	125	77083	50	30833	75	46250
Сумма			**1597500**		**945000**		**652500**
Среднее			**266250**		**157500**		**108750**

Cov(x, v) равна 157500 и Cov(x, w) = 108750.

Мы проверили, что Cov(x, y) = Cov(x, v)+ Cov(x, w).

- Именно так и должно быть. Рассмотрим i - ю семью
- Поскольку
- $y_i = v_i + w_i$ и

$$\bar{y} = \bar{v} + \bar{w}.$$

$$(x_i - \bar{x})(y_i - \bar{y}) = (x_i - \bar{x})(v_i + w_i - \bar{v} - \bar{w}) = (x_i - \bar{x})(v_i - \bar{v}) + (x_i - \bar{x})(w_i - \bar{w})$$

Таким образом, вклад семьи i в Cov(x, y) является суммой ее вкладов в Cov(x, v) и Cov(x, w).

Тоже самое справедливо для всех семей и, соответственно, для ковариации в целом.

- Правило 2:
- Если y = a z, где a - константа,

то Cov(x, y) = a Cov(x, z).

Семья	Доход семьи (x)	Расходы на питание и одежду (y)	Вторая выборка: расходы семьи на питание и одежду (z)
1	3000	1100	2200
2	2500	850	1700
3	4000	1200	2400
4	6000	1600	3200
5	3300	1000	2000
6	4500	1300	2600

- Последняя колонка (z) дает расходы на питание и одежду для второго множества из 6 семей.
- Каждое наблюдение z=2y.
- Предполагается, что значения величины x для второго набора семей являются такими же, как и ранее.

Семья	$(x-\bar{x})$	$(z-\bar{z})$	$(x-\bar{x})(z-\bar{z})$
1	-883	-150	132500
2	-1383	-650	899167
3	117	50	5833
4	2117	850	1700167
5	-583	-350	204167
6	617	250	154167
Сумма			**3195000**
Среднее			**532500**

Из таблицы можно видеть, что **Cov(x, z)** равна **532500**, что равно **2Cov(x, y)**

Таким образом мы проверили, что **Cov(x, 2y) = 2Cov(x, y).**

$$Cov(x,y) = 2Cov(x,y) = \frac{1}{n}\sum(x_i-\bar{x})(z_i-\bar{z}) = \frac{1}{n}\sum(x_i-\bar{x})(ay_i-\bar{y}) = \frac{a}{n}\sum(x_i-\bar{x})(y_i-\bar{y}) = aCov(x,y)$$

- **Правило 3**:
- Если y = a, где a - константа,

то **Cov(x, y) = 0**.

Допустим, что каждая семья в выборке имеет по два взрослых человека, и предположим, что по недоразумению вы решили вычислить ковариацию между общим доходом (x) и числом взрослых в семье (a).

Естественно, что $a_1=a_2...=a_6$ =2= среднему значению.

Поэтому Cov(x, a)=0.

Семья	x	a	$(x - \bar{x})$	$(a - \bar{a})$	$(x - \bar{x})(a - \bar{a})$
1	3000	2	-883	0	0
2	2500	2	-1383	0	0
3	4000	2	117	0	0
4	6000	2	2117	0	0
5	3300	2	-583	0	0
6	4500	2	617	0	0
Сумма	**23300**	**12**			**0**
Среднее	**3883**	**2**			**0**

Выборочная дисперсия, правила расчета дисперсии

- Для выборки из n наблюдений $x_1, \ldots, x_n$ выборочная дисперсия определяется как среднеквадратичное отклонение в выборке:

$$Var(x) = \frac{1}{n}\sum (x_i - \bar{x})^2$$

Ранее была определена исправленная", или несмещенная, выборочная дисперсия :

$$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

- Заметим, что дисперсия переменной x может рассматриваться как ковариация между двумя величинами x:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = Cov(x,x),$$

Кроме того можно получить другую формулу:

$$Var(x) = [\frac{1}{n}\sum_{i=1}^{n}(x_i)^2] - \bar{x}^2$$

- Существует несколько правил для расчета дисперсии, которые являются аналогами правил для ковариации.
- **Правило 1**: Если y = v + w,

то Var(y) = Var(v) + Var(w) + 2Cov(v, w).

- Доказательство :

Если y = v + w, то

Var(y) = Cov(y, y) = Cov(y, [v + w]) =

= Cov([v + w], v) + Cov([v + w], w), по правилу ковариации 1,

= Cov(v, v) + Cov(w, v) + Cov(v, w) + Cov(w, w), по правилу ковариации 1,

= Var(v) + Var(w) + 2Cov(v, w).

- **Правило 2**: Если y = a z, где a - константа,

то $Var(y) = a^2 Var(z)$.

- Доказательство:

Дважды используя правило ковариации 2, получим:

$Var(y) = Cov(y, y) = Cov(y, az) = a\ Cov(y, z) =$

$= a\ Cov(az, z) = a^2\ Cov(z, z) = a^2 Var(z)$.

- **Правило 3**: Если y = a, где a - константа, то Var(y) = 0.
- По правилу ковариации 3 имеем:

Var(y) = Cov(a, a) = 0

- Действительно, если **y** - постоянная, то ее среднее значение является той же самой постоянной и равняется нулю для всех наблюдений.
- Следовательно, **Var(y)=0.**

- **Правило 4**: Если y = v + a, где a - константа, то **Var(y) = Var(v).**
- Доказательство:
- Если y = v + a, где a - константа, то по правилу ковариации 1, используя затем правила 1 и 3 для дисперсии и правило 3 для ковариации, получаем:

Var(y) = Var(v + a) = Var(v) + Var(a) + 2Cov(v, a) = Var(v).

Корреляционная зависимость

- **Функциональная зависимость-** связь, при которой каждому значению независимой переменной х значение переменной у
- **Статистическая зависимость** – связь, при которой каждому значению независимой переменной х соответствует множество значений зависимой переменной у , причем неизвестно заранее, какое именно значение у.

- Частным случаем статистической зависимости является **корреляционная зависимость.**
- **Корреляционная зависимость-** связь, при которой **каждому значению независимой переменной соответствует определенное математическое ожидание (среднее значение) независимой переменной.**

- Корреляционная связь является «неполной» зависимостью, которая проявляется не в каждом отдельном случае, а только в средних величинах при достаточно большом числе случаев.
- Корреляционная зависимость исследуется с помощью методов корреляционного и регрессионного анализа.

- Наиболее разработанной в эконометрике является методология **парной линейной регрессии**, рассматривающая влияние переменной **x** на переменную **y** и представляющая собой **однофакторный корреляционный и регрессионный анализ**.

Корреляционный анализ

- Заключается в количественном определении тесноты связи между двумя признаками (при парной связи) и между результативным и множеством факторных признаков (рои многофакторной связи)
- **Корреляция** – это статистическая зависимость между случайными величинами, при которой **изменение одной из случайных величин приводит к изменению математического ожидания другой.**

Коэффициент корреляции

- **Коэффициент корреляции** является более точной мерой зависимости между величинами.
- Подобно дисперсии и ковариации, **коэффициент корреляции** имеет две формы - теоретическую и выборочную.
- Теоретический коэффициент корреляции ρ для переменных x и y определяется следующим образом:

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sqrt{\sigma_x^2 \quad \sigma_y^2}}$$

- Если x и y независимы, то $\mathbf{p_{x,y}}$ **=0**, так как равна нулю теоретическая ковариация.
- Если между переменными существует положительная зависимость, то теоретический **коэффициент корреляции будет положительным**.
- Если существует строгая положительная зависимость, то он примет максимальное значение, равное **1**.
- Аналогичным образом при отрицательной зависимости теоретический коэффициент **корреляции будет отрицательным с минимальным значением -1**.

Качественные характеристики связи

коэфф. корреляции	*вид связи*
от 0 *до* $\lvert\pm 0{,}3\rvert$	*отсутствует*
от $\lvert\pm 0{,}3\rvert$ *до* $\lvert\pm 0{,}5\rvert$	*слабая*
от $\lvert\pm 0{,}5\rvert$ *до* $\lvert\pm 0{,}7\rvert$	*умеренная*
от $\lvert\pm 0{,}7\rvert$ *до* $\lvert\pm 1{,}0\rvert$	*сильная*

- Выборочный коэффициент корреляции **r** для переменных x и y определяется путем замены теоретических дисперсий и ковариации в формуле теоретического коэффициента корреляции на их несмещенные оценки:

$$r_{x,y} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

- Выборочный коэффициент корреляции имеет максимальное значение, равное 1, которое получается при строгой линейной положительной зависимости между выборочными значениями x и y, и минимальное значение -1, когда существует линейная отрицательная зависимость.
- **Величина r=0 показывает, что зависимость между наблюдениями x и y в выборке отсутствует, но это не говорит о том, что p=0, и наоборот.**

- Рассмотрим пример расчета корреляции.
- Уже вычислена **Cov(p, y)= -16,24**, поэтому необходимы вычислить только **Var(p) и Var(y).**

Наблюдение	p	y	$(p-\bar{p})$	$(y-\bar{y})$	$(p-\bar{p})^2$	$(y-\bar{y})^2$
1	103,5	26,2	-39,86	-0,07	1588,82	0,01
2	127,0	24,8	-16,36	-1,47	267,65	2,16
3	126,0	25,6	-17,36	-0,67	301,37	0,45
4	124,8	26,8	-18,56	0,53	344,47	0,28
5	124,7	27,7	-18,66	1,43	348,20	2,05
6	121,6	28,3	-21,76	2,03	473,50	4,12
7	149,7	27,4	6,34	1,13	40,20	1,28
8	188,8	25,1	45,44	-1,17	2064,79	1,37
9	193,6	25,2	50,24	-1,07	2524,06	1,15
10	173,9	25,6	30,54	-0,67	932,69	0,45
Сумма	**1433,6**	**262,7**			**8885,75**	**13,30**
Среднее	**143,36**	**26,27**			**888,58**	**1,33**

В последних двух колонках таблицы можно найти, что Var(p) составляет 888,58 и Var(y) равна 1,33.

$$r = \frac{-16{,}24}{\sqrt{888{,}58 \times 1{,}33}} = \frac{-16{,}24}{34{,}38} = -0{,}47$$

- Из примера видим, что коэффициент корреляции незначительно отличается от нуля.
- Одна из причин в получении такого результата заключается в очень небольшом размере выборки.

- Еще одна причина - не **учтено влияние увеличения дохода на потребительский спрос в целом и на спрос на бензин в частности.**
- **Положительный эффект увеличения дохода в основном компенсировал отрицательный эффект роста цен, и, таким образом, спрос на бензин оставался стабильным**.

- Чтобы выделить эти два фактора используют коэффициент частной корреляции:

$$r_{xyz} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{[1 - r_{xz}^2][1 - r_{yz}^2]}}$$

где $r_{xy.z}$ - коэффициент частной корреляции между **x** и **y** в случае постоянства воздействия величины **z**, а r_{xy}, r_{xz} **и** r_{yz} - обычные коэффициенты корреляции между x и y, x и z, y и z соответственно.

- В примере со спросом на бензин можно вычислить корреляцию между ценой и располагаемым личным доходом и между спросом и доходом.
- Результаты по данной выборке составят соответственно 0,84 и 0,02.
- Подставим результаты в уравнение частной корреляции.

$$r = \frac{-0,47 - 0,84 \cdot 0,02}{\sqrt{(1 - 0,84^2) \cdot (1 - 0,02^2)}} = -0,91$$

Результат получился лучше

Выводы

- Таким образом, корреляция может быть 3-х видов:
  - Парная – связь между двумя признаками
  - Частная – зависимость между двумя признаками при фиксированном значении других признаков.
  - Множественная – зависимость результативным признаком и двумя и более факторными признаками.

- Коэффициенты корреляции как статистические величины подвергаются в анализе оценке на достоверность
- Для оценки значимости коэффициента корреляции используется t- критерий Стьюденте.

- Выдвигается гипотеза о равенстве нулю коэффициента корреляции r_{xy} =0.
- Если гипотеза отвергается, то коэффициент корреляции признается значимым, а связь между переменными существенной.

Формула расчета критерия Стьюдента

$$t_{расч} = r_{xy}\sqrt{\frac{n-k-1}{1-r_{xy}^2}},$$

где k – число факторов в модели

- Значение t критерия сравнивают с табличным (n-k-1 число степеней свободы, уровень значимости обычно 0,05 или 0,1)
- Если $t_{расч} > t_{табл}$, то значение коэффициента корреляции признается значимым, делается вывод что между исследуемыми переменными есть тесная статистическая взаимосвязь.