МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ГЕОГРАФИИ

Карпиченко Александр Александрович

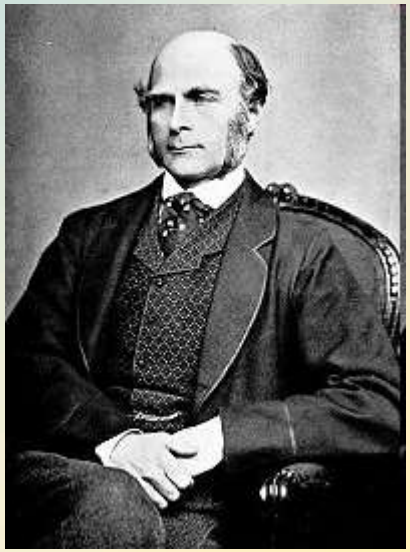
доцент кафедры почвоведения и земельных информационных систем

Литература

- elib.bsu.by
- Математические методы в географии: учебно-методическое пособие / Н. К. Чертко, А. А. Карпиченко. – Минск: БГУ, 2009.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Основоположником теории корреляции считаются английские биометрики Френсис Гальтон (1822–1911) и Карл Пирсон (1857–1936). Термин «корреляция» означает соотношение, соответствие. Представление о корреляции как о взаимозависимости случайных переменных величин лежит в основе статистической теории корреляции – изучение зависимости вариации признака от окружающих условий.



5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Одни признаки выступают в роли *влияющих (факторных)*, другие – на которые влияют, *результативных*. Зависимости между признаками могут быть *функциональными и корреляционными*. Функциональные связи характеризуются полным соответствием между изменением факторного признака и изменением результативной величины. Каждому значению признака-фактора соответствует определенное значение результативного признака.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

В корреляционных связях между изменением факторного и результативного признака нет полного соответствия. В сложном взаимодействии находится сам результативный признак. Поэтому результаты корреляционного анализа имеют значение в данной связи, а интерпретация этих результатов в общем виде требует построения системы корреляционных связей. Они характеризуются множеством причин и следствий и с их помощью устанавливается тенденция изменения результативного признака при изменении величины факторного признака. Например, на производительность труда влияют факторы степени совершенствования техники и технологии, уровень механизации и автоматизации труда, специализации производства, текучесть кадров и т. д.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

В природе и обществе явления и события протекают по характеру корреляционной связи, когда при изменении величины одного признака существует тенденция изменения другого признака. Корреляционная связь – это частный случай статистической связи. ***Корреляционный анализ используется при установлении тесноты зависимости между явлениями, процессами, объектами***.

Целью исследования часто бывает установление взаимосвязи (корреляции) между признаками. Знание зависимости дает возможность решать кардинальную задачу любого исследования – возможность предвидеть, прогнозировать развитие ситуации при изменении влияющего фактора. **С помощью корреляции можно дать лишь формальную оценку взаимосвязей**. Поэтому прежде чем приступать к вычислению коэффициентов корреляции между любыми признаками, следует теоретически установить, имеется ли между этими признаками взаимосвязь. Ведь формально статистика может доказать несуществующие связи, например, между высотой здания в городе и урожайностью пшеницы в фермерских хозяйствах.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Связь между явлениями (корреляция) определяется путем постановки опытов, статистического анализа. **Корреляцию не следует отождествлять с причинностью**. Необходимо иметь в виду, что доказательство математической связи должно опираться на реальную зависимость между явлениями.

Любой показатель связи служит приближенной оценкой рассматриваемой зависимости и не является гарантией существования жесткой (функциональной) соподчиненности. Отсутствие жесткой зависимости в природе и обществе способствует саморегуляции процессов, явлений, систем.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

По направлению связь может быть *прямой* и *обратной*; **по характеру** – *функциональной* или *статистической (корреляционной)*; **по величине** – *слабой, средней* или *сильной*; **по форме** – *линейной* и *нелинейной*; **по количеству коррелируемых признаков** – *парной и множественной*.

Функциональная зависимость характерна для геометрических форм, технических систем, когда каждому значению одного признака соответствует точное значение другого. Это пример взаимосвязи площади прямоугольника и длины его одной из сторон. Такая зависимость полная или исчерпывающая.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Выделяют несколько видов парной корреляционной связи:

- параллельно-соотносительную, или ассоциативную, когда оба признака изменяются сопряжено, частично под действием общих причин и следствий (приуроченность растительности и почв к определенным формам рельефа; развития промышленности и рост населения к сырьевым ресурсам);
- субпричинную, когда один фактор выступает как отдельная причина сопряженного изменения признака (связь биомассы с количеством осадков; рост населения и рождаемости);
- взаимоупреждающую, когда причина и следствие, находясь в устойчивой взаимной связи, последовательно влияют друг на друга (влажность воздуха и осадки).

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Если на признак влияет несколько факторов, то приходится оценивать множественную корреляцию. *Множественная корреляция* служит основой выявления связей между признаками, но требует строгой нормальности и прямолинейности распределения, поэтому использование ее может быть затруднено. С ростом числа переменных объем вычислительных работ увеличивается пропорционально квадрату числа переменных. В этом случае труднее оценивать значимость результатов, так как увеличиваются ошибки коэффициентов корреляции. Практически в таких случаях ограничиваются изучением лишь главных факторов. Однако характер влияния главных факторов на признак более детально и точно исследуют путем факторного анализа.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

В практической работе по установлению корреляции между признаками и явлениями необходимо придерживаться следующей последовательности:

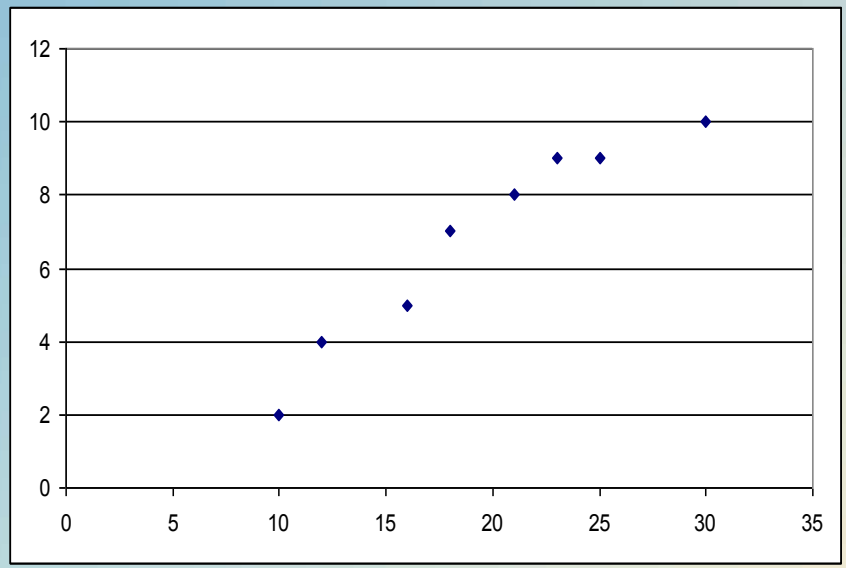
- на основании проведенных исследований предварительно определяют, существует ли связь между рассматриваемыми признаками;
- если связь между ними существует, устанавливают ее форму, направление и тесноту, используя график.

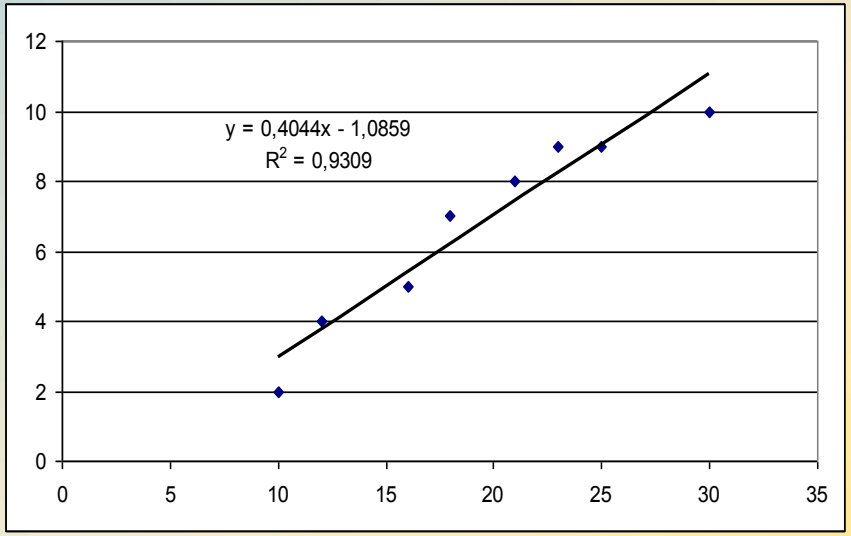
В начале составляются сопряженные вариационные ряды, в которых следует определить аргумент *x* и функцию *y*:

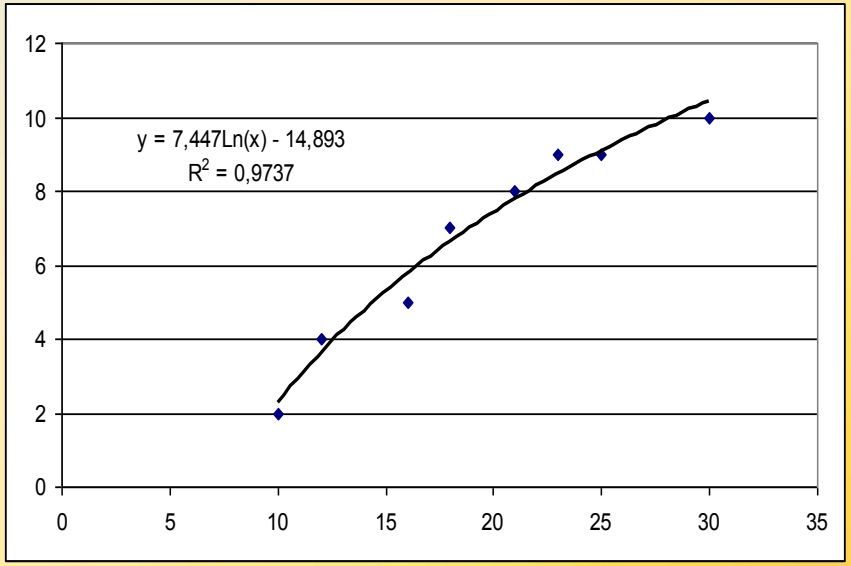
x	10	12	16	18	21	23	25	30
y	2	4	5	7	8	9	9	10

По сопряженным вариантам строится график, который помогает установить вид зависимости между аргументом и функцией. От формы корреляционной связи зависит дальнейшая обработка экспериментальных или статистических данных.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ







Линейная зависимость предполагает вычисление коэффициента корреляции ***r***, а нелинейная – корреляционного отношения **η** (эта).

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

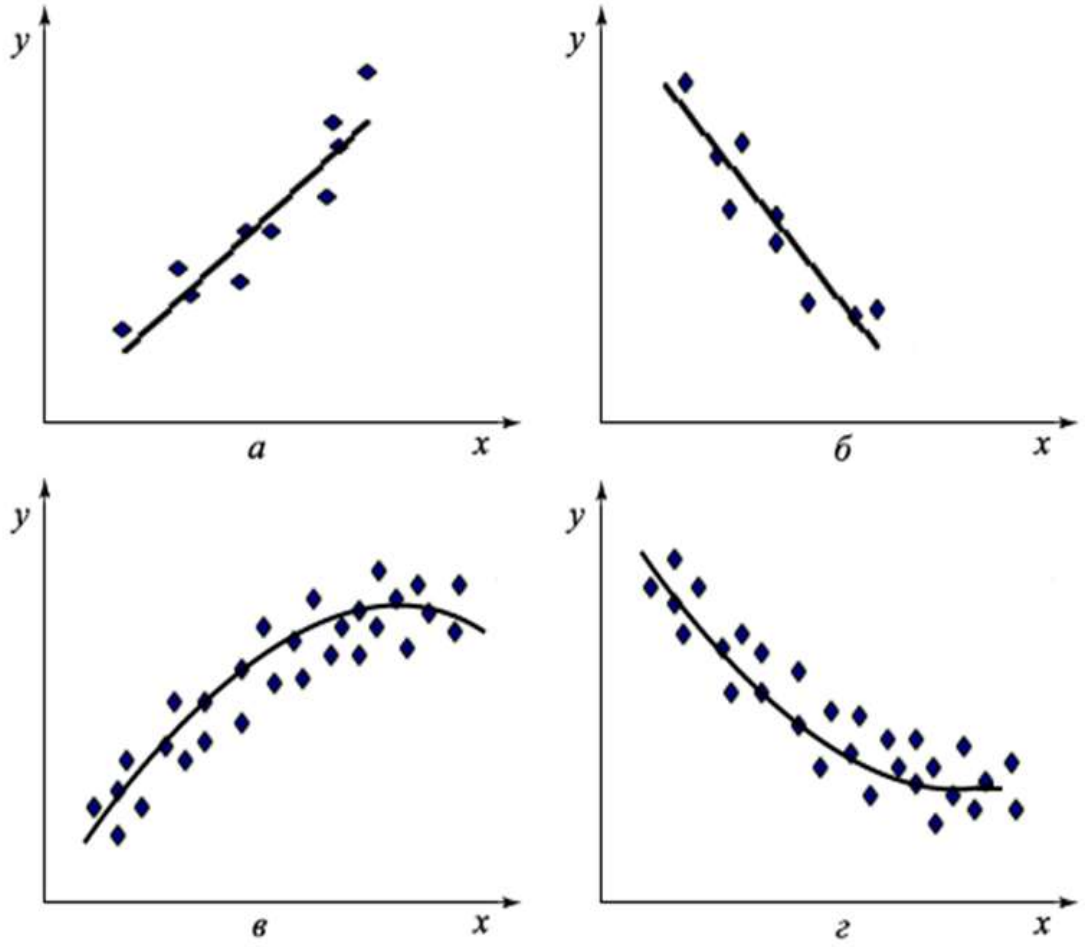


Рис. 5.1. Форма корреляционной связи:
а – прямая линейная; *б* – обратная линейная; *в* – парабалическая; *г* – гиперболическая

Линейная зависимость предполагает вычисление коэффициента корреляции ***r***, а нелинейная – корреляционного отношения **η**.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

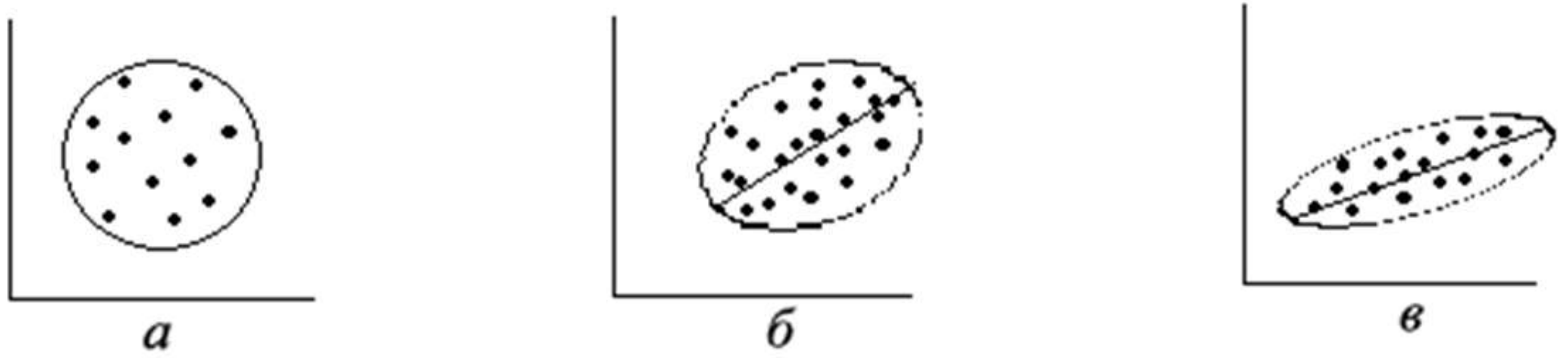


Рис. 5.2. Степень рассеяния частот и величина связи:
а – $r \approx 0$; *б* – $r \approx 0{,}5$; *в* – $r \approx 0{,}7$

Степень рассеяния частот или вариант относительно линии регрессии на графике указывает ориентировочно на тесноту связи: чем меньше рассеяние, тем сильнее связь.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Корреляционный анализ решает следующие задачи:

- установление направления и формы связи;
- оценка тесноты связи;
- оценка репрезентативности статистических оценок взаимосвязи;
- определение величины детерминации (доли взаимовлияния) коррелируемых факторов.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Для оценки связи используют следующие численные критерии (коэффициенты) корреляционной связи:

- коэффициент корреляции (r) при линейной зависимости,
- корреляционное отношение (η) при нелинейной зависимости,
- коэффициенты множественной регрессии,
- ранговые коэффициенты линейной корреляции Спирмена или Кендэла.

5.1. Линейная корреляция

Линия регрессии по координатам точек на графике проводится таким образом, чтобы точки в равном количестве находились по обе стороны линии.

Более точное значение *r* получаем расчетным способом следующим образом как при прямой (*r* от 0 до 1), так и при обратной (*r* от 0 до –1) зависимости:

$$r = \frac{\sum (x_i - M_x)(y_i - M_y)}{\sqrt{\sum (x_i - M_x)^2 \sum (y_i - M_y)^2}},$$

где $(x_i - M_x)$, $(y_i - M_y)$ – отклонения значений индивидуальных вариант x_i и y_i от их средних значений M_x и M_y.

5.1. Линейная корреляция

Принимается следующая характеристика тесноты корреляционной связи: если r (η) = 0 ± 0,4, то связь считается слабой; от ± 0,4 до ± 0,7 – средняя; от ± 0,7 до ± 1 – сильная; r = ± 1 и η = 1 – связь считается функциональная.

Достоверность вычисленного коэффициента корреляции может быть установлена двумя путями: путем сравнения с табличным значением r (прил. 7 учебника); второй путь – через критерий Стьюдента. Если $r_{выч} > r_{табл}$, то влияние фактора на признак достоверно; если меньше табличного – не достоверно.

5.1. Линейная корреляция

Достоверность связи устанавливается путем сравнения *r* (η) расчетного (фактического) и *r* (η) теоретического (табличного). Если $r(\eta)_{выч} > r(\eta)_{табл}$ при учете степени свободы (v) вариационных рядов и уровня вероятности P = 0,95 и 0,99, то связь между признаками доказана без учета величины *r*(η).

Регрессионный анализ обычно является продолжением корреляционного в случае если *r* (η) ≥ ± 0,7.

5.1. Линейная корреляция

При использовании критерия Стьюдента для доказательства достоверности *r* вначале рассчитывают стандартную ошибку коэффициента корреляции:

$$m_r = \sqrt{(1 - r^2)/(N_n - 2)}, \qquad (5.3)$$

где N_n – число сопряженных пар в сравниваемых выборках.

Значение коэффициента корреляции записывают с учетом его ошибки и уровня значимости: $r_{0,95\ (0,99)} \pm m_r$. Затем вычисляют критерий Стьюдента для коэффициента корреляции:

$$t_r = r / m_r \qquad (5.4)$$

Критерий Стьюдента можно рассчитать иначе:

$$t_r = r\sqrt{N_n - 2} / \sqrt{1 - r^2} \qquad (5.5)$$

5.1. Линейная корреляция

Пример. Исследованиями установлено, что на содержание подвижного марганца в почве влияет реакция среды. Необходимо доказать достоверность установленной зависимости. Получены следующие исходные данные (х – гидролитическая кислотность, мэкв на 100 г почвы; у – содержание подвижного марганца, мг/кг почвы):

x	83	72	69	90	90	95	95	91	75	70
y	56	42	18	84	56	107	90	58	31	48

5.1. Линейная корреляция

Исходные данные для расчета коэффициента корреляции

x_i	$x_i - M_x$	$(x_i - M_x)^2$	y_i	$y_i - M_y$	$(y_i - M_y)^2$	$(x_i - M_x) \cdot (y_i - M_y)$
69	–14	196	18	–42	1764	558
70	–13	169	48	–12	144	156
72	–11	121	42	–18	324	198
75	–8	64	31	–29	841	232
83	0	0	56	–4	16	0
90	7	49	84	24	576	168
90	7	49	56	–4	16	–28
91	8	64	68	8	64	64
95	12	144	90	30	900	360
95	12	144	107	47	2209	564
$\sum$ 830 M_x 83	$\sum$ 0	$\sum$ 1000	$\sum$ 600 M_y 60	$\sum$ 0	$\sum$ 6854	$\sum$ 2302

5.1. Линейная корреляция

Достоверность связи устанавливается путем сравнения *r* (η) расчетного (фактического) и *r* (η) теоретического (табличного). Если $r(\eta)_{выч} > r(\eta)_{табл}$ при учете степени свободы (v) вариационных рядов и уровня вероятности P = 0,95 и 0,99, то зависимость между признаками доказана без учета величины *r*(η). Регрессионный анализ обычно является продолжением корреляционного в случае если *r* (η) ≥ ± 0,7.

Коэффициент детерминации (*причинности*) R^2 (D^2) – это коэффициент корреляции, возведенный в квадрат, например, $R^2 = r^2 = 0{,}2^2 = 0{,}04$. С помощью коэффициента детерминации можно установить долю влияния анализируемого факторного признака на результативный признак. В случае, когда $R^2 = 0{,}04$, можно утверждать, что доля влияющего фактора (*x*) на признак (*y*) составляет 4 %. Следовательно, на долю других факторов приходится 96 % влияния.

5.2. Нелинейная корреляция

Зависимость между признаками не всегда выражается в виде прямой линии. Если рассеяние точек на графике приближается к кривой линии, то зависимость устанавливается с использованием корреляционного отношения (η), величина которого изменяется только от 0 до 1. Для него теоретические значения приводятся отдельно в таблице или находятся при перерасчете его в критерий Стъюдента. При нелинейной корреляции вычисляется корреляционное отношение (η).

Для установления формы связи иногда используется критерий криволинейности в случаях, когда кривая линия мало отличается от прямой. Существует несколько способов оценки степени криволинейности.

5.2. Нелинейная корреляция

Первый способ менее точный. Оценка степени криволинейности определяется по разности коэффициента корреляции и корреляционного отношения использованием неравенства: $\eta^2 - r^2 \geq 0{,}1$. Корреляция считается криволинейной, если полученный результат соответствует этому неравенству. Предварительно следует рассчитать между сравниваемыми выборками r и η.

Второй способ оценки степени криволинейности связан с применением критерия Стъюдента:

$$t = 0{,}5\sqrt{\frac{N}{(\eta^2 - r^2)^{-1} - 2 + (\eta^2 + r^2)}} \geq 3.$$

Если $t_{выч} < 3$ или $t_{выч} < t_{табл}$, то рассматриваемая связь несущественно отклоняется от прямолинейной, поэтому относим ее к линейной. В других случаях связь между признаками относят к криволинейной и рассчитывается корреляционное отношение.

Корреляционное отношение, как и коэффициент корреляции, используется для оценки прямой и обратной зависимости между признаками.

5.2. Нелинейная корреляция

Оценка прямой нелинейной зависимости между признаками. Прямая нелинейная зависимость определяется как **параболическая**. Расчет корреляционного отношения производится по формуле с использованием функции *y*:

$$\eta = \sqrt{\frac{n\sum(\bar{y} - M_y)^2}{\sum(y_i - M_y)^2}}, \qquad (5.6)$$

где $\bar{y}$ – среднее арифметическое частных групп по y_i; n – число вариант в частной группе; $\bar{y} - M_y$ – отклонение общего среднего (M_y) от средних арифметических частных групп ($\bar{y}$).

Ошибка корреляционного отношения независимо от способа расчета вычисляется следующим образом:

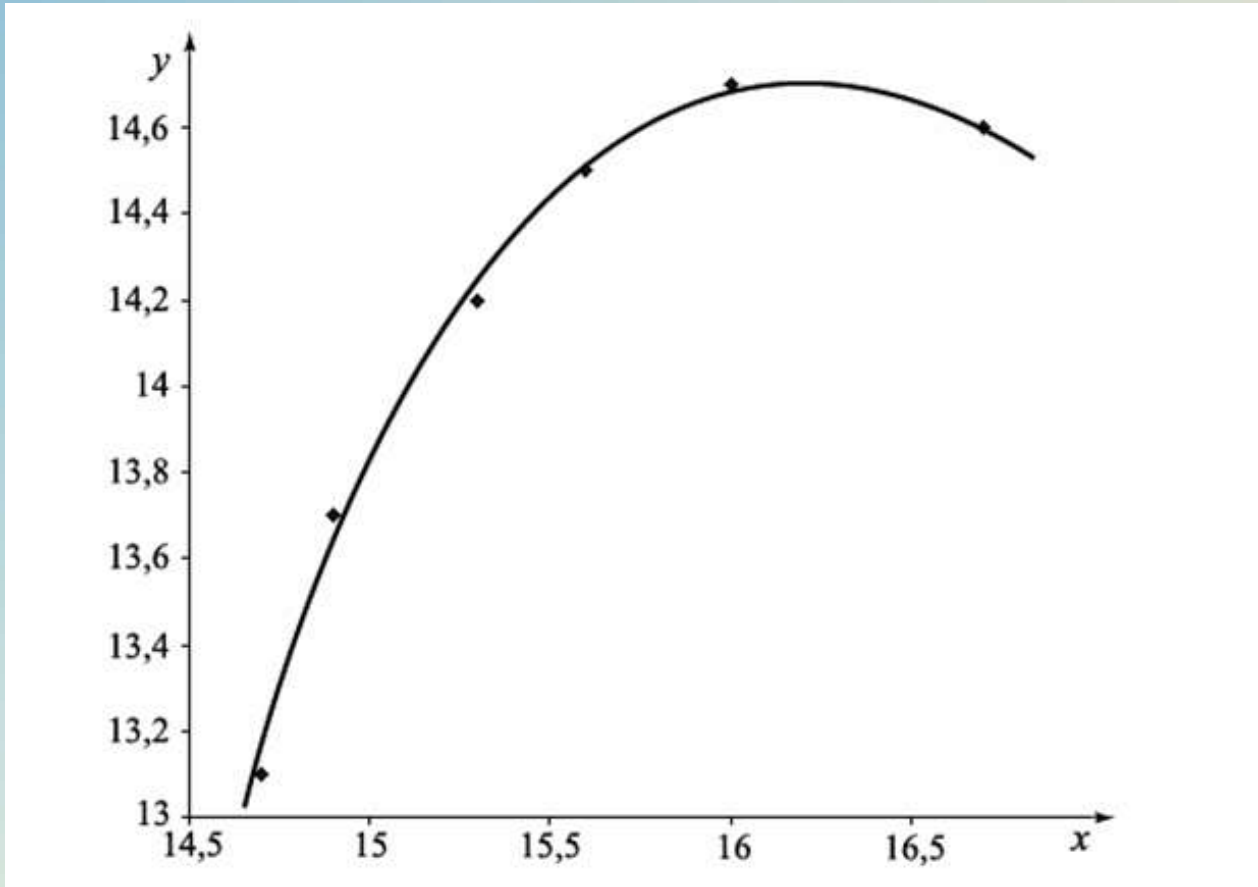
$$m_\eta = \sqrt{(1-\eta^2)/(N_{нар} - 2)} \qquad (5.7)$$

5.2. Нелинейная корреляция

Следует установить, существует ли зависимость между температурой воздуха (x, ^oC) и упругостью водяного пара (y, мбар) по шести метеорологическим постам Беларуси исходя из следующих данных:

x_i	14,7	14,9	15,3	15,6	16,0	16,7
y_i	13,3	13,7	14,2	14,5	14,7	14,6

5.2. Нелинейная корреляция



Для данной корреляции t_{η} = 2,51 < $t_{табл}$= 2,78 при *P* 0,95 для v = 4, поэтому значение корреляционного отношения следует признать не доказанным, а зависимость между температурой воздуха и упругостью водяного пара положительна, но не достоверна.

5.2. Нелинейная корреляция

Оценка обратной нелинейной зависимости между признаками. Алгоритм вычисления и доказательств при расчете корреляционного отношения обратной нелинейной (гиперболической) зависимости аналогичен алгоритму прямой нелинейной зависимости. Различие состоит в том, что в качестве исходных вариант используется выборка со значениями *x*.

Для нелинейной обратной (гиперболической) зависимости корреляционное отношение определяется с использованием аргумента *x* по формуле, условные обозначения в которой аналогичны формуле параболической зависимости:

$$\eta_x = \sqrt{\frac{n\sum(\bar{x}_{гр} - M_x)^2}{\sum(x_i - M_x)^2}}$$

5.2. Нелинейная корреляция

Расчетные величины η по *x* сравнивают с табличными (теоретическими) для степени свободы (v = $N_{пар}$ – 1) и *P* = 0,95 и 0,99. Если расчетная величина больше табличной, то можно утверждать с уверенностью о наличии достоверной зависимости между признаком и фактором.

Для всех коэффициентов можно рассчитать их ошибки: $r \pm m_r$; $\eta \pm m_n$.

При расчете η с использованием выборочных вариант *x* и *y* можно также применить следующие формулы с известными обозначениями:

$$\eta_x = \sqrt{\frac{\sum (x_i - M_x)^2 - (x_i - \bar{x}_{гр})^2}{\sum (x_i - M_x)^2}}$$

$$\eta_y = \sqrt{\frac{\sum (y_i - M_y)^2 - (y_i - \bar{y}_{гр})^2}{\sum (y_i - M_y)^2}}$$

5.4. Множественная корреляция

Метод множественной корреляции применяется в случаях, когда необходимо установить совокупное влияние всего комплекса факторов на результативный признак. Величина коэффициента множественной корреляции изменяется от 0 до 1. Его можно вычислить с использованием коэффициентов частной линейной корреляции по формуле:

$$R_{1,23} = \sqrt{1-(1-r_{12}^2)(1-r_{13}^2)} = \sqrt{1-(0{,}4^2)(1-(-0{,}7)^2)} = 0{,}75$$

По коэффициенту R = 0,75 определяется коэффициент детерминации R^2 (R_D) = $0{,}75^2$ = 0,56. Он показывает, что доля совместного влияния второго и третьего признаков составляет 56 %.

5.6. Ранговая корреляция

В географических исследованиях иногда приходится обрабатывать быстро и с наименьшими затратами фактический материал, даже если получаются менее точные результаты. В некоторых случаях работают с качественной информацией или с громоздкими вычислениями. В таких случаях для установления зависимости между признаками используется ранговая корреляция.

Процесс упорядочения вариант по какому-либо признаку (например, увеличение или уменьшение количества населения по районам) называют ранжированием. Каждому члену ранжированного ряда присваивается *ранг.* Для обозначения рангов, как правило, используются числа в пределах единиц и десятков, например: 1, 2, 3, ..., *n.* Первой варианте или группе вариант присваивается ранг 1, второй варианте или группе – 2 и т.д. Следует иметь в виду, что одни и те же варианты в зависимости от цели группировки могут иметь различные ранги. Величина ранга не позволяет нам судить о том, насколько близко друг к другу расположены на шкале измерения различные варианты совокупности или качественные признаки.

5.6. Ранговая корреляция

Ранговую корреляцию можно применять для всех упорядоченных признаков (например, экспертные оценки, баллы, бонитеты). Объем сопряженных выборок должен быть не менее пяти.

Для расчета зависимости (x, y) существуют следующие коэффициенты ранговой корреляции: коэффициент неупорядоченности $r_н$ и коэффициент Спирмена $r_с$.

5.6. Ранговая корреляция

Коэффициент ранговой корреляции Спирмена рассчитать легче, чем коэффициент неупорядоченности, поэтому в естественных науках предпочтение отдается r_c. Коэффициент Спирмена представляет собой следующее соотношение:

$$r_c = 1 - \frac{6\Sigma(x'-y')^2}{N_n^3 - N_n}, \text{ или } r_c = 1 - \frac{6\Sigma(d^2)}{N_n^3 - N_n},$$

где d – разность между сопряженными рангами; x' – величины рангов, заменяющие фактические варианты или качественные признаки по аргументу x; y' – величины рангов, заменяющие фактические варианты или качественные признаки по функции y; $N_п$ – количество сопряженных пар.

Достоверность полученного рангового коэффициента можно установить аналогично достоверности коэффициента корреляции.