Серыя «У дапамогу педагогу» заснавана ў 1995 годзе

Навукова-метадычны часопіс
Выдаецца з IV квартала 1995 года
Пасведчанне аб дзяржаўнай рэгістрацыі сродку масавай інфармацыі
№ 641 ад 04.09.2009 г., выдадзенае Міністэрствам інфармацыі Рэспублікі Беларусь
Выходзіць штомесячна з II паўгоддзя 2012 года

# Геаграфія



Заснавальнік і выдавец —
Рэспубліканскае ўнітарнае прадпрыемства «Выдавецтва «Адукацыя і выхаванне»
Міністэрства адукацыі Рэспублікі Беларусь

Вул. Будзённага, 21, 220070, г. Мінск;
тэл.: 297-93-24 (адк. сакратар), 297-93-22 (аддзел маркетынгу),
факс: 297-91-49, e-mail: geography@aiv.by, http://www.aiv.by

**3(148) сакавік 2018**

# ЗМЕСТ






*...тыкі, а паводле магчымасцей камп'ютара.*

...бор, макет і вёрстка *В. Ю. Лагун.*

...рук афсетны. Папера афсетная.

...арк. 7,44.

...вабодная.

...адказнасцю «СУГАРТ»,

... корп. 2, каб. 287, 220012, г. Мінск.

*А. С. Соколов,*
*старший преподаватель кафедры экологии*
*Гомельского государственного университета им. Ф. Скорины*

# МЕТОДЫ ОЦЕНКИ СВЯЗИ МЕЖДУ ПОКАЗАТЕЛЯМИ

*Окончание. Начало в № 2*

Изучение природных объектов, обладающих большим количеством различных числовых характеристик, позволяет поставить перед исследователем вопрос о наличии связей (*корреляций*) между различными характеристиками. Например, между средней температурой за сутки и уровнем активности каких-либо насекомых, между уровнем грунтовых вод и проективным покрытием мха кукушкин лён, между плотностью населения на определённых территориях и значением лесистости этих территорий. Корреляция может быть положительной (когда возрастание одного показателя сопровождается возрастанием второго) или отрицательной (когда наоборот). Во многих работах эта связь декларируется умозрительно, без привлечения статистического аппарата, что не даёт оснований утверждать о её объективном существовании. Применение статистических методов позволит 1) определить наличие связи, т. е. изменяются ли два сравниваемых параметра независимо друг от друга (нулевая гипотеза) или взаимосвязано (альтернативная гипотеза); 2) вычислить количественное значение тесноты этой связи.

Как и в других разделах, здесь могут использоваться методы *параметрические* (условием применения которых являются *нормальные распределения* сопоставляемых переменных) и *непараметрические* (не имеющие этого ограничения).

Наиболее популярным показателем является **коэффициент корреляции Пирсона** — метод параметрической статистики, позволяющий определить наличие или отсутствие линейной связи между двумя количественными показателями, а также оценить её тесноту и статистическую значимость. Другими словами, критерий корреляции Пирсона даёт возможность определить, есть ли линейная связь между изменениями значений двух переменных. В статистических расчётах и выводах коэффициент корреляции обычно обозначается как $r_{xy}$ или $R_{xy}$.

При корреляционном анализе необходимо помнить следующее:

- понятия «связь» и «зависимость» не тождественны. Используя коэффициент Пирсона, мы можем лишь утверждать, что связь есть, но мы не можем утверждать, что эта связь имеет причинно-следственный характер и что один из параметров зависит от другого. В этом заключается одна из основных ошибок исследователей, которые, получив результат, говорящий о наличии корреляционной статистической связи между двумя параметрами, интерпретируют его так, что один параметр зависит от другого. Они могут оба зависеть от какого-то третьего фактора;
- коэффициент Пирсона может определить наличие только ***линейной*** связи. Соответственно, низкое значение коэффициента Пирсона или статисти-

ческая незначимость рассчитанного коэффициента говорит не о том, что связи нет ВООБЩЕ, а о том, что нет ЛИНЕЙНОЙ связи, хотя какая-нибудь другая связь вполне может быть;

- с помощью коэффициента Пирсона можно определить только наличие и тесноту связи. Прочие характеристики связи, в том числе направление (прямая или обратная), характер изменений (прямолинейный или криволинейный), а также наличие зависимости одной переменной от другой определяются при помощи *регрессионного анализа*;
- коэффициент Пирсона может определить связь только между двумя показателями. Если необходимо определить наличие взаимосвязей между тремя и более показателями, пользуются методами *факторного анализа*;
- значения коэффициента Пирсона могут быть от −1 до +1. Эти два значения свидетельствуют об абсолютной связи, и в реальности в естественнонаучных исследованиях практически не встречаются. Значение 0 означает полное отсутствие связи; значение коэффициента менее 0,3 свидетельствуют о *слабой* связи, 0,3 до 0,7 — о связи *средней* тесноты, более 0,7 — о *сильной* связи.

Коэффициент корреляции Пирсона вычисляется по формуле:

$$r_{xy} = \frac{\sum\left(d_x \cdot d_y\right)}{\sqrt{\left(\sum d_x^2 \cdot \sum d_y^2\right)}},$$

где $d_x$ — отклонение значения каждого показателя в вариационном ряду $x$ от среднего арифметического для этого ряда; $d_y$ — отклонение значения каждого показателя в вариационном ряду $y$ от среднего арифметического для этого ряда ($x$ и $y$ — это ряды данных, между которыми мы проверяем наличие связи). При расчёте коэффициента Пирсона удобно сначала заполнить таблицу следующего вида (в качестве примера проверим связь между проективным покрытием мха кукушкин лён, % ($x$) и уровнем грунтовых вод, м ($y$) (табл. 5).

*Таблица 5* — Пример для применения коэффициента корреляции Пирсона

| $x$ | $y$ | $d_x$ | $d_y$ | $d_x^2$ | $d_y^2$ | $d_x \cdot d_y$ |
|---|---|---|---|---|---|---|
| 65 | 0,2 | −9,75 | 0,29 | 95,06 | 0,08 | −2,80 |
| 23 | 1,3 | 32,25 | −0,81 | 1040,06 | 0,66 | −26,20 |
| 88 | 0,1 | −32,75 | 0,39 | 1072,56 | 0,15 | −12,69 |
| 45 | 0,6 | 10,25 | −0,11 | 105,06 | 0,01 | −1,15 |
| 27 | 1,1 | 28,25 | −0,61 | 798,06 | 0,38 | −17,30 |
| 59 | 0,4 | −3,75 | 0,09 | 14,06 | 0,01 | −0,33 |
| 73 | 0,1 | −17,75 | 0,39 | 315,06 | 0,15 | −6,88 |
| 62 | 0,1 | −6,75 | 0,39 | 45,56 | 0,15 | −2,62 |
| *Ср. = 55,3* | *Ср. = 0,49* | | | *Σ = 3485,50* | *Σ = 1,59* | *Σ = −69,98* |

Таким образом, подставляя полученные значения в формулу, вычисляем коэффициент корреляции Пирсона:

$$r_{xy} = \frac{-68,98}{\sqrt{(3485,50 \cdot 1,59)}} = -0,93.$$

Таким образом, связь между уровнем грунтовых вод и проективным покрытием мха кукушкин лён оказалась отрицательной и очень тесной.

Осталось проверить статистическую значимость полученного значения. Она осуществляется при помощи $t$-критерия, рассчитываемого по следующей формуле:

$$t_r = \frac{\left|r_{xy}\right| \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}.$$

В нашем примере

$$t_r = \frac{0,93 \cdot \sqrt{8-2}}{\sqrt{1-0,86}} = 6,09.$$

Число степеней свободы равно $f = n - 2$, то есть в нашем примере 6. По таблице критических значений $t$-критерия находим критическое значение для этого числа степеней свободы. Видим, что полученное нами значение $t$ превышает критическое как для $p < 0{,}05$, так и для $p < 0{,}01$, а значит, наличие связи между рассматриваемыми характеристиками считается статистически доказанным, т. е. связь является статистически значимой при $p < 0{,}01$.

**Коэффициент Спирмена** применяется, когда зависимость между двумя показателями не является линейной или когда распределение параметров, между которыми проверяется наличие связи, не является нормальным. Также он применяется, когда неизвестны числовые значения характеристик показателей, а известны лишь их ранги. Поэтому его ещё одно название — *коэффициент ранговой корреляции*.

Свойства данного коэффициента такие же, как у коэффициента Пирсона — изменяется от −1 до +1, тем больше абсолютное значение, чем больше теснота связи.

Для расчёта необходимо для каждого абсолютного значения рассматриваемых параметров вычислить значение его ранга. Затем разности рангов каждой пары возвести в квадрат, суммировать полученные результаты и рассчитать коэффициент Спирмена по формуле

$$\rho_{xy} = 1 - \frac{6 \cdot \sum d^2}{n \cdot \left(n^2 - 1\right)},$$

где $d$ — разность рангов каждой пары параметров. Для примера рассчитаем коэффициент Спирмена по тому же примеру, по которому рассчитывали коэффициент Пирсона (табл. 6).

*Таблица 6* — Пример для применения коэффициента корреляции Симпсона

| *x* | *y* | ранг *x* | ранг *y* | *d* | $d^2$ |
|---|---|---|---|---|---|
| 65 | 0,2 | 6 | 4 | 2 | 4 |
| 23 | 1,3 | 1 | 8 | −7 | 49 |
| 88 | 0,1 | 8 | 2 | 6 | 36 |
| 45 | 0,6 | 3 | 6 | −3 | 9 |
| 27 | 1,1 | 2 | 7 | −5 | 25 |
| 59 | 0,4 | 4 | 5 | −1 | 1 |
| 73 | 0,1 | 7 | 2 | 5 | 25 |
| 62 | 0,1 | 5 | 2 | 3 | 9 |
| | | | | | Σ = 158 |

$$\rho_{xy} = 1 - \frac{6 \cdot 158}{8 \cdot \left(8^2 - 1\right)} = -0{,}88.$$

Теперь необходимо проверить статистическую значимость полученного результата. Это осуществляется точно так же, как для прошлой формулы — по значению $t$-критерия, только вместо $r_{xy}$ в формулу подставлять вычисленное значение коэффициента Спирмена $\rho_{xy}$. Число степеней свободы тоже определяется аналогично. В данном случае $t$-критерий будет равен 6,22, а значит, проверив по таблице критических значений, убеждаемся, что корреляция между этими показателями статистически значима. При расчёте коэффициента желательно, чтобы значения показателей каждого параметра не повторялись, это обеспечит более точный результат.

Кроме параметров, характеризуемых количественно, в природных системах существуют и параметры, которые характеризуются качественно — типы почв, экологические группы организмов, поло-

жение в пределах речных бассейнов, наличие или отсутствие определённых организмов или признаков в сообществе и т. д. Между такими параметрами тоже может существовать статистическая связь. Рассмотрим статистические показатели, которые показывают связь между двумя параметрами, характеризующимися лишь такими признаками, как «наличие» или «отсутствие». К примеру, необходимо определить, есть ли связь какого-либо вида растения с определённым типом почв — то есть растёт ли этот вид преимущественно на этой почве или никакой связи нет. Для этого необходима совокупность наблюдений, в каждом из которых фиксировалось бы наличие или отсутствие данного вида растений, а также наличие или отсутствие данного типа почв. В результате, все наблюдения ($N$) можно будет разделить на 4 группы:

$A$ — наблюдения, где фиксировался и вид растений, и тип почвы;

$B$ — наблюдения, в которых фиксировался только данный вид растений, а данный тип почв отсутствовал;

$C$ — наблюдения, в которых фиксировался только данный тип почв, а данный вид растений отсутствовал;

$D$ — наблюдения, где не были отмечены ни данный вид, ни данная почва.

Одним из наиболее популярных статистических методов, с помощью которых проверяют наличие связи между двумя качественными показателями, является **критерий хи-квадрат Пирсона.**

При применении критерия хи-квадрат полученные данные наблюдений заносятся в таблицу, которая называется четырёхпольная таблица сопряжённости (табл. 7).

*Таблица 7*

| | Параметр I есть (1) | Параметр I нет (0) | Всего |
|---|---|---|---|
| Параметр II есть (1) | $A$ | $B$ | $A + B$ |
| Параметр II отсутствует (0) | $C$ | $D$ | $C + D$ |
| Всего | $A + C$ | $B + D$ | $A + B + C + D$ |

Таблица может состоять и более чем из 4 полей, а каждый параметр может оцениваться более чем двумя качественными состояниями, однако такие методы не затрагиваются в настоящей статье.

При расчёте критерия хи-квадрат необходимо сначала *рассчитать ожидаемое количество наблюдений* для каждой из ячеек таблицы сопряжённости (то есть, как распределились бы наблюдения при условии справедливости нулевой гипотезы об отсутствии взаимосвязи) путём перемножения сумм рядов и столбцов с последующим делением полученного произведения на общее число наблюдений (табл. 8).

*Таблица 8*

| | Параметр I есть (1) | Параметр I нет (0) | Всего |
|---|---|---|---|
| Параметр II есть (1) | $(A + B)*(A + C)/(A + B + C + D)$ | $(A + B)*(B + D)/(A + B + C + D)$ | $A + B$ |
| Параметр II отсутствует (0) | $(C + D)*(A + C)/(A + B + C + D)$ | $(C + D)*(B + D)/(A + B + C + D)$ | $C + D$ |
| Всего | $A + C$ | $B + D$ | $A + B + C + D$ |

Затем полученные для каждого из 4 полей таблицы значения подставляем в формулу для расчёта критерия хи-квадрат:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)}{E_{ij}},$$

где $i$ — номер строки (от 1 до $r$), $j$ — номер столбца (от 1 до c), $O_{ij}$ — фактическое количество наблюдений в ячейке $ij$, $E_{ij}$ — рассчитанное ожидаемое число наблюдений в ячейке $ij$.

Число степеней свободы равно: $f = (r - 1) \cdot (c - 1)$. То есть для четырёхпольной таблицы, в которой 2 ряда ($r = 2$) и 2 столбца ($c = 2$), число степеней свободы составляет всегда 1 ($f_{2x2} = (2 - 1) \cdot (2 - 1) = 1$).

*Таблица 9* — Критические значения критерия $\chi^2$

| $f$ | $p < 0{,}05$ | $p < 0{,}01$ | $f$ | $p < 0{,}05$ | $p < 0{,}01$ | $f$ | $p < 0{,}05$ | $p < 0{,}01$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 3,841 | 6,635 | 4 | 9,488 | 13,277 | 7 | 14,067 | 18,475 |
| 2 | 5,991 | 9,21 | 5 | 11,07 | 15,086 | 8 | 15,507 | 20,09 |
| 3 | 7,815 | 11,345 | 6 | 12,592 | 16,812 | 9 | 16,919 | 21,666 |

Осталось только сравнить полученное значение $\chi^2$ с критическим значением по таблице 9 при числе степеней свободы, равном 1.

Если рассчётное значение выше критического, то нулевая гипотеза отвергается, и наличие связи между двумя параметрами считается статистически доказанным.

Для определения количественного значения тесноты этой связи, её знака предлагается ряд коэффициентов, из которых можно выделить трансформированный коэффициент Дайса:

$$K_D = \frac{A - \min(B,C)}{A + \min(B,C)},$$

где $A$ — число наблюдений в ячейке таблицы $A$ (то есть число наблюдений, где присутствовали оба параметра), $\min(B, C)$ — минимальное значение из ячеек $B$ и $C$. Коэффициент может принимать значения от −1 (когда нет ни одного случая совместной встречи двух параметров) до +1 (когда по крайней мере один из двух параметров никогда не встречался без второго).

В том случае, если число ожидаемого явления меньше 10 хотя бы в одной ячейке, при анализе четырёхпольных таблиц должен рассчитываться **критерий хи-квадрат с поправкой Йейтса.** Данная поправка позволяет уменьшить вероятность ошибки первого типа, т. е. обнаружения различий там, где их нет. Формула для расчёта критерия хи-квадрат с поправкой Йейтса следующая:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(\left|O_{ij} - E_{ij}\right| - 0{,}5\right)^2}{E_{ij}}.$$

К примеру, необходимо выяснить взаимосвязь в распространении двух видов растений — крапивы двудомной и недотроги обыкновенной. Для этого были использованы описания 132 пробных площадей. Из них на 58 встречались оба вида, на 22 — только крапива на 38 — только недотрога. Таким образом, $N$ = 132, $A$ = 88, $B$ = 22, $C$ = 12, $D$ = 10.

Рассчитав ожидаемое количество наблюдений, получаем (табл. 10).

*Талбца 10*

| | Крапива есть (1) | Крапивы нет (0) | Всего |
|---|---|---|---|
| Недотрога есть (1) | (88 + 22)*(88 + 12)/ (88 + 22 + 12 + 10) = **83,3** | (88 + 22)*(22 + 10)/ (88 + 22 + 12 + 10) = **26,7** | $A + B$ |
| Недотроги нет (0) | (12 + 10)*(88 + 12)/ (88 + 22 + 12 + 10) = **16,7** | (12 + 10)*(22 + 10)/ (88 + 22 + 12 + 10) = **5,3** | $C + D$ |
| Всего | $A + C$ | $B + D$ | $A+B+C+D$ |

Поскольку в ячейке $D$ значение получилось меньше 10, то вычислим хи-квадрат с поправкой Йейтса:

$$\chi^2 = \frac{(|88-83,3|-0,5)^2}{83,3} + \frac{(|22-26,7|-0,5)^2}{26,7} + \frac{(|12-16,7|-0,5)^2}{16,7} + \frac{(|10-5,3|-0,5)2}{5,3} = 5,26.$$

Поскольку 5,26 больше, чем критическое значение 3,84, то считается доказанным, что крапива и недотрога распространены не независимо, между ними существует взаимосвязь, а значит, они обладают сходной реакцией на те факторы и условия природной среды, которые были представлены на исследованных пробных площадях. Значение $K_D = (88 - 12) / (88 + 12) = +0,76$.

Если хотя бы в одной ячейке ожидаемое явление меньше 5, то для анализа должен использоваться **точный критерий Фишера**. Он вычисляется по формуле:

$$P = \frac{(A+B)! \cdot (C+D)! \cdot (A+C)! \cdot (B+D)!}{A! \cdot B! \cdot C! \cdot D! \cdot N!},$$

где $N$ — общее число наблюдений; ! — знак факториала, представляющего собой произведение числа на последовательность чисел, каждое из которых меньше предыдущего на 1 (например, $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$).

Отличительной особенностью точного критерия Фишера является то, что полученное значение $P$ уже является уровнем значимости, при котором отвергается нулевая гипотеза и принимается альтернативная. Если это значение меньше 0,05, то признаётся доказанное наличие статистической связи между двумя показателями.

К примеру, изучалась взаимосвязь распространения ослинника двулетнего и песчаных почв. Всего было сделано 85 описаний, из них 15 описаний, где отмечены и песчаная почва, и ослинник, 2 описания, где отмечен ослинник, но почва не песчаная, 40 описаний, где почва песчаная, но нет ослинника и 28 описаний, где не отмечено ни песчаной почвы, ни ослинника. Таким образом, $N = 85$, $A = 15$, $B = 2$, $C = 40$, $D = 28$. Вычислим точный критерий Фишера:

$$P = \frac{(15+2)! \cdot (40+28)! \cdot (15+40)! \cdot (2+28)!}{15! \cdot 2! \cdot 40! \cdot 28! \cdot 86!} = 0,016,$$

Полученное значение 0,016 меньше 0,05, т. е. связь между песчаными почвами и ослинником двулетним считается доказанной.

## Список использованной литературы


1. *Атраментова, Л. А.* Статистические методы в биологии: учебник для студ. высш. уч. зав. / Л. А. Атраментова, О. М. Утевская. — Горловка : Лихтар, 2008. — 248 с.
2. *Филандышева, Л. Б.* Статистические методы в географии: учеб.-метод. пособие / Л. Б. Филандышева, Е. С. Сапьян; отв. ред. А. В. Пучкин. — Томск : Издательский Дом Томского государственного университета, 2015. — 164 с.
3. *Любимов, В. Б.* Математические методы в экологии: учеб. пособие / В. Б. Любимов, И. В. Мельников, А. В. Силенок. — Брянск : РИО БГУ, 2017. — 201 с.
4. *Крючков, А. В.* Биометрия: учеб. пособие / А. В. Крючков, И. В. Маракулин. — Киров : Изд-во ВятГУ, 2011. — 87 с.