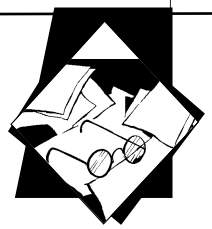


Методические разработки и рекомендации

В разделе публикуются методики и рекомендации, имеющие как общеметодологический, так и узкопредметный характер. Материалы этого раздела призваны помочь в практической организации учебного исследования самому широкому кругу воспитателей — как профессиональным педагогам (и школ, и учреждений дополнительного образования), так и родителям.

Школа начинающего исследователя

Продолжаем занятия в школе начинающего исследователя природы. На занятиях вы познакомитесь с основами работы в полевых условиях, а также правилам и анализа и представления исследовательского материала. Вам будет предложена возможность знакомства с различными

Использование математических методов в биологических исследованиях школьников[1]

Хайтов Вадим Михайлович,
кандидат биологических наук, заведующий лабораторией экологии морского бентоса (гидробиологии) Санкт-Петербургского городского Дворца творчества юных, г. Санкт-Петербург

Описание взаимосвязи величин

В этой главе речь в пойдет о *корреляционном анализе*. В практике биологических исследований очень часто встает вопрос о том, имеется ли какая-нибудь взаимосвязь между явлениями. Здесь нам надо немного отойти от чистой статистики и чуть-чуть поговорить о философии.

[1] Продолжение. Начало: ИРШ. 2008. №1. С. 42–57; С. 56–66.

Всем хорошо известно, что существует так называемая причинно-следственная связь между явлениями. Например, связь между ветром и деревьями. Это связь причинно-следственная. Ветер дует — деревья качаются. Ветер является причиной раскачивания деревьев. Но вспомните себя в глубоком детстве! Описанное соотношение было для вас абсолютно неочевидным — вполне вероятным казалось, что раскачивание деревьев и есть причина возникновения ветра.

Не надо думать, что подобные перевороты в сознании происходят только с детьми, когда те взрослеют, в аналогичную ситуацию много раз попадали и большие ученые. Часто причинно-следственные связи, которые воспринимались одним образом, позднее начинают восприниматься иначе. Но при всех подобных изменениях в понимании процессов само наличие связи между явлениями не подвергается сомнению. Сколько бы ни было вам лет, связь между ветром и раскачиванием деревьев останется. Вот это и называется *корреляцией*.

Корреляция — это не что иное, как наличие взаимосвязи между явлениями. Причем, заметьте: ничего о причинах и следствиях здесь не говорится. Наличие корреляции не означает, что между явлениями есть причинно-следственная связь, но какая-то связь есть. Она может быть чисто случайной, а может и нет. **Поиск причин этой связи не входит в компетенцию статистики!** Статистика занимается лишь ее **выявлением**. Корреляционный же анализ призван количественно измерить силу этой взаимосвязи.

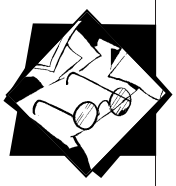
Разберем простейший вариант корреляционного анализа, применяемый для анализа связи между явлениями, которые регистрируются в виде количественных данных. Для этого вернемся к примеру с рачками-бокоплавами Pontoporeia femorata, который мы разобрали ранее (см. № 2 за 2008 год). У каждого рачка мы изучали два параметра — длину тела и вес. Оба этих параметра были измерены количественно. Мы можем поставить вопрос о наличии связи между этими величинами. Для ответа на поставленный вопрос надо провести корреляционный анализ. Для простоты возьмем не всю выборку, а лишь первые 29 значений.

Таблица 17. **Параметры тела (длина и вес) рачков-бокоплавов Pontoporeia femorata**

L	*P*		*L*	*P*
5,7	4		4,7	3
6,6	10		5,1	4
7,3	11		5,2	7
4,7	5		5,8	6
5,5	6		4,7	4
5,4	5		5,6	6
6,2	10		4,3	3
6,3	11		5,8	7
10,5	32		4,4	2
13,5	56		4,3	5
6,3	10		9,2	22
5,9	7		4,2	3
6,2	6		4,4	3
5,8	6		6,7	8
4,3	4			

методиками изучения природы и будут даваться задания по их освоению. Материалы занятий будут полезны не только школьникам, но и педагогам, желающим организовать исследовательскую деятельность в своих школах, школьных лесничествах, кружках и клубах.

Ведет занятия Николай Харитонов, заведующий отделом Московского городского программно-методического центра дополнительного образования детей (nikol-2@yandex.ru).



Биология — наука, использующая различные методы представления данных: записи наблюдений, рисунки, численные данные (результаты учетов, измерений, взвешиваний и т.п.). В предыдущих номера хречь шла о математических методах в биологических исследованиях, представленных статистическими методами; давалось описание методов сравнения двух величин и методов анализа структуры популяции. В этом номере представлено описание взаимосвязи величин. В следующем номере будут представлены методы многомерного анализа.

Для измерения силы взаимосвязи между количественными параметрами применяется *коэффициент корреляции Браве–Пирсона* (*r*). Этот коэффициент может принимать значения от –1 до 1. Если $r = 1$ или $r = -1$, то корреляция очень сильная — явления очень сильно взаимосвязаны. Если $r = 0$, то явления не связаны друг с другом. Огромное значение имеет и знак коэффициента корреляции. Если коэффициент положительный, то это означает, что чем больше выражено первое явление, тем больше выражено второе. Если коэффициент отрицательный, то связь обратная: чем больше выражено первое явление, тем меньше выражено второе. Вычисление этого коэффициента производится по следующей формуле:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}.$$

В этой формуле x_i, y_i — значения каждого отдельного измерения признака X и Y у *i*-го объекта, — средние значения признаков X и Y, соответственно. Эту формулу можно записать несколько иначе — в виде, более удобном для вычисления:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{N}}{\sqrt{(\sum x_i^2 - \frac{(\sum x_i)^2}{N})(\sum y_i^2 - \frac{(\sum y_i)^2}{N})}}.$$

Для вычисления по этой формуле надо рассчитать входящие в нее величины.

Подставляем полученные величины в формулу (1). Итак, мы получили достаточно высокий коэффициент корреляции. Однако мы должны еще проверить, достоверно ли это значение. Ведь мы изучили не всех бокоплавов на Земле, а лишь небольшую выборку. Для ответа на вопрос о достоверности коэффициента корреляции проще всего использовать так называемую таблицу пороговых значений (см. табл. 3).

Таблица 18. **Ход вычисления коэффициента корреляции Браве—Пирсона**

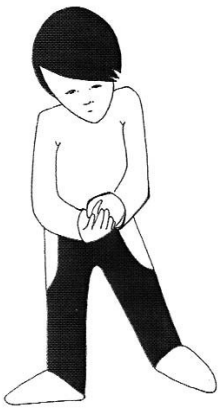
	L	P
Сумма $\sum L_i$ (*или* $\sum P_i$)	174,6	266,0
Квадрат суммы $(\sum L_i)^2$ (*или* $(\sum P_i)^2$)	30485,2	70756,0
Сумма квадратов $\sum L_i^2$ (*или* $\sum P_i^2$)	1165,3	5756,0
Сумма произведений $\sum L_i P_i$	2193,10	
Объем выборки N	29	

$$r = \frac{2193{,}10 - \frac{174{,}6 \cdot 266{,}0}{29}}{\sqrt{\left(1165{,}3 - \frac{30485{,}2}{29}\right)\left(5756.0 - \frac{70756{,}0}{29}\right)}} = 0{,}96. \qquad (1)$$

Для работы с ней нам надо знать, как обычно, доверительную вероятность и число степеней свободы. Первую величину мы принимаем равной $P_{\text{дов}} = 95\ \%$. Число степеней свободы вычисляется по следующей формуле:

$$n = N - 2,$$

где N — число пар наблюдений. При $n = N - 2 = 29 - 2 = 27$ и $P_{\text{дов}} = 95\ \%$ пороговое значение коэффициента корреляции будет равно $r_{\text{порог}} = 0{,}367$. Эмпирическое значение коэффициента (здесь необходимо брать коэффициент по модулю, т.е. без учета знака) значительно превышает пороговое. Таким образом, вычисленная корреляция в высшей степени достоверна, а значит мы можем с уверенностью говорить, что между длиной тела бокоплава и его весом связь существует. Более того, мы можем однозначно утверждать, что чем больше длина тела животного, тем больше его вес.



Внимание, контрольный вопрос. А можем ли мы на основании проведенных исследований утверждать, что увеличение длины тела рачка является причиной увеличения его веса? Конечно же, нет! Ведь мы вычислили лишь корреляцию, то есть, только установили наличие связи, но не провели изучение причинно-следственных отношений. Изучение этого аспекта — это уже задача биологии, а не статистики.

Итак, мы рассмотрели методы изучения взаимосвязи между явлениями, которые описываются количественными данными. Но у приведенного метода есть одно ограничение о котором нельзя забывать. Этот метод применим не для всех данных, которые выражены численно. Он применим лишь для тех величин, которые имеют **нормальное распределение** (см. начало статьи в № 1 за 2008 год). Проверка нормальности распределения величины — задача достаточно сложная, и здесь мы ее рассматривать не будем. Однако многие начинающие биологи по умолчанию считают, что большинство признаков биологических объектов, которые можно численно измерить, подчиняются закону нормального распределения. Часто это не так. Например, не подчиняются нормальному распределению всевозможные доли (проценты), зачастую, временные промежутки и расстояния на местности. Как же быть в тех случаях, когда вы не уверены в том, что изученная величина подчиняется закону нормального распределения? Для решения задач, в которых используются подобные величины, лучше пользоваться методами непараметрической статистики. В случае корреляционного анализа используются *ранговые коэффициенты корреляции.* Самый распространенный метод анализа в таких случаях был

предложен Спирменом (поэтому коэффициент, предложенный им, обычно называют *коэффициентом Спирмена* (r_S)).

Рассмотрим применение такого коэффициента на следующем примере. Предположим, вы задались целью выяснить, существует ли взаимосвязь между следующими явлениями: степенью зараженности прудовиков партенитами печеночного сосальщика и близостью водоема, в котором обитают прудовики, к деревенским пастбищам. Первое явление вы оценили с помощью доли зараженных моллюсков в некоторой выборке, взятой в том или ином водоеме. Вторую величину вы оценивали как расстояние от водоема до ближайшего пастбища. Все полученные данные вы свели в следующую таблицу:

***Таблица 19.* Удаленность водоема от пастбища и доля зараженных прудовиков**

№ водоема	Расстояние до пастбища (м)	Доля загрязненных прудовиков (%)
1	0	80
2	12	20
3	11	20
4	0	60
5	1000	5
6	300	2
7	10	40
8	10	20
9	25	50
10	33	20
11	80	20
12	500	10
13	255	10
14	100	20
15	121	22
16	0	65

Несмотря на то что в нашей таблице приведены количественные данные, взять и вычислить коэффициент корреляции Браве–Пирсона нельзя. Это связано с тем, что ни расстояния, ни доли зараженных моллюсков не имеют нормального распределения. Для выявления корреляции в данном случае необходимо произвести ранжирование данных. Эта процедура очень простая, но требует достаточно внимательной работы (особенно при обработке больших массивов данных).

Итак, давайте осуществим процедуру ранжирования. Для начала все значения в каждом изученном признаке (параметре) надо упорядочить по мере возрастания. Разберем этот первый

шаг на примере данных, оценивающих расстояние от пруда до пастбища:

Таблица 20. **Упорядоченные данные по расстоянию (первый шаг ранжирования)**

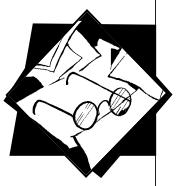
№ водоема	Расстояние до пастбища	Порядковый номер
1	0	1
4	0	2
16	0	3
7	10	4
8	10	5
3	11	6
2	12	7
9	25	8
10	33	9
11	80	10
14	100	11
15	121	12
13	255	13
6	300	14
12	500	15
5	1000	16

Далее каждому баллу присваивается его порядковый номер в ряду, упорядоченном по мере возрастания значения признака:

Таблица 21. **Присвоение порядкового номера значениям расстояния (второй шаг ранжирования)**

№ водоема	Расстояние до пастбища	Ранги для значений расстояния
1	0	2
4	0	2
16	0	2
7	10	4,5
8	10	4,5
3	11	6
2	12	7
9	25	8
10	33	9
11	80	10
14	100	11
15	121	12
13	255	13
6	300	14
12	500	15
5	1000	16

Эти порядковые номера и были бы рангами, если бы среди них не было бы повторяющихся значений. Так, например, водоемы 1, 4 и 16 имеют нулевое расстояние от пастбища, но им



присвоены разные порядковые номера. Аналогично, водоемы 7 и 8 расположены на одинаковом расстоянии. В такой ситуации всем объектам, имеющим одинаковые значения признака, присваивается ранг, равный среднему значению их порядковых номеров[1].

Таблица 22. **Ранги для значений расстояний (заключительный этап ранжирования)**

№ водоема	Расстояние до пастбища	Порядковый номер
1	0	1
4	0	2
16	0	3
7	10	4
8	10	5
3	11	6
2	12	7
9	25	8
10	33	9
11	80	10
14	100	11
15	121	12
13	255	13
6	300	14
12	500	15
5	1000	16

Проведя аналогичные операции с величинами, описывающими зараженность моллюсков, переписываем исходную таблицу.

Таблица 23. **Ранговые оценки расстояния от водоема до пастбища и зараженности прудовиков**

№ водоема	Расстояние до пастбища	Доля зараженных прудовиков	Ранги для значений расстояния	Ранги для значений зараженности
1	0	80	2	16
2	12	20	7	7,5
3	11	20	6	7,5
4	0	60	2	14
5	1000	5	16	2
6	300	2	14	1
7	10	40	4,5	12
8	10	20	4,5	7,5
9	25	50	8	13
10	33	20	9	7,5
11	80	20	10	7,5
12	500	10	15	3,5
13	255	10	13	3,5
14	100	20	11	7,5
15	121	22	12	11
16	0	65	2	15

1 Крайне желательно спланировать сбор материала так, чтобы повторяющихся значений было меньше, в противном случае придется вычислять некоторые поправки, суть которых мы в данном пособии обсуждать не будем.

Внимание! Если вы по каким-то причинам выкинули некоторые пары наблюдений или, наоборот, добавили их, то всю процедуру ранжирования надо повторить сначала!

Теперь можно приступить к вычислениям коэффициента ранговой корреляции Спирмена. Он рассчитывается по следующей формуле:

$$r_s = 1 - \frac{6 \cdot \sum (x - y)^2}{N^3 - N}.$$

В этой формуле x — ранг первого признака в паре, y — ранг второго признака в паре, N — число пар.

Для удобства вычислений построим такую следующую таблицу.

Таблица 24. **Ход вычисления коэффициента Спирмена**

Ранги для значений расстояния	Ранги для значений зараженности	x-y	$(x-y)^2$
2	16	-14	196
7	7,5	-0,5	0,25
6	7,5	-1,5	2,25
2	14	-12	144
16	2	14	196
14	1	13	169
4,5	12	-7,5	56,25
4,5	7,5	-3	9
8	13	-5	25
9	7,5	1,5	2,25
10	7,5	2,5	6,25
15	3,5	11,5	132,25
13	3,5	9,5	90,25
11	7,5	3,5	12,25
12	11	1	1
2	15	-13	169
Сумма $(x-y)^2$			1211

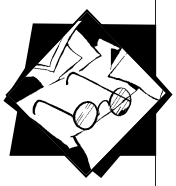
N=16

N^3=4096

Теперь можно подставить все значения в формулу:

$$r_s = 1 - \frac{6 \cdot 1211}{4096 - 16} = -0{,}781.$$

Полученное значение коэффициента корреляции далее необходимо сравнить с пороговым значением ($r_{s\text{-порог}}$), найденным по специальной таблице (табл. 4). Эта таблица очень похожа на таблицу пороговых значений коэффициента корреляции, которую мы использовали ранее (табл. 3 в № 2 за 2008 год). Однако,



поскольку мы работали не истинными значениями признаков, а с их рангами, то появились небольшие отличия. Для $n = 16 - 2 = 14$ при $P_{\text{дов}} = 95$ % оно равно $r_{\text{s-порог}} = 0{,}501$. Поскольку полученное эмпирическое значение (без учета знака) выше табличного, мы можем с высокой степенью уверенности утверждать, что два изученных явления взаимосвязаны. Более того, отрицательное значение коэффициента корреляции говорит о том, что чем дальше от пастбища расположен пруд, тем меньше доля зараженных моллюсков. Очевидно, что здесь имеет место некоторая биологическая закономерность, которая лежит на поверхности (прудовики являются переносчиками сосальщиков). Однако описание этой закономерности — задача не статистическая. С помощью статистики мы лишь выявили наличие связи и доказали, что она значима (неслучайна).

Теперь перейдем к разговору об измерении силы взаимосвязи между явлениями, которые характеризуются качественными данными. Напомним, что это данные, принимающие только два значения (0 или 1, да или нет, + или –), то есть, явление есть или явления нет.

Предположим, что вы решили выяснить, существует ли взаимосвязь между наличием тараканов в квартире и ежедневными уборками в ней. Вооружившись блокнотом, вы обошли всех своих друзей и у каждого из них выясняли следующую информацию: есть ли у них в квартире тараканы и делается ли уборка ежедневно. Будучи уже грамотным исследователем, вы оформили свои наблюдения в виде таблицы.

Таблица 25. **Наличие тараканов и ежедневная уборка («+» — явление присутствует, «–» — явление не присутствует)**

№ опроса	Наличие тараканов	Наличие ежедневной уборки
1	-	+
2	+	+
3	+	-
4	-	-
5	-	+
6	+	-
7	-	+
8	+	+
9	-	+
10	-	+
11	+	+
12	+	-
13	+	+
14	-	+
15	+	+
16	+	-
17	+	-
18	-	-
19	+	-
20	-	-

После того как первичные результаты наблюдений оформлены, можно приступить к изучению взаимосвязи. Для этого составляют так называемую *четырехпольную таблицу*. Эта таблица имеет следующий вид.

Таблица 26. **Вид четырехпольной таблицы**

		Явление I	
		«+»	«-»
Явление II	«+»	a	b
	«-»	c	d

В этой таблице a — число наблюдений, в которых присутствуют оба явления; b —число наблюдений, в которых имеет место второе явление, но отсутствует первое; c — число наблюдений, в которых присутствует первое явление, но отсутствует второе; d — число наблюдений, в которых не проявилось ни первое, ни второе явление.

Для нашего случая с тараканами мы можем построить следующую четырехпольную таблицу.

Таблица 27. **Четырехпольная таблица для выяснения взаимосвязи уборки и наличия тараканов**

		Наличие ежедневной уборки	
		«+»	«-»
Наличие тараканов	«+»	5	6
	«-»	6	3

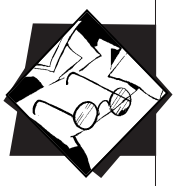
Теперь необходимо вычислить соответствующий показатель согласованности изменений для качественных признаков — так называемый *коэффициент ассоциации* (r_a). Этот коэффициент, как и коэффициент корреляции, отражает силу связи между явлениями. Он также может принимать значения от –1 до 1. Если $r_a = 1$, то явления сцеплены друг с другом — если происходит одно, то происходит и другое. Если $r_a = -1$, то при проявлении одного явления не проявляется другое. Если $r_a = 0$, то явления не связаны друг с другом.

В разбираемом нами случае положительная корреляция означала бы, что ежедневная уборка увеличивает возможность встретить тараканов, а наличие отрицательной — напротив, означает, что проведение ежедневной уборки уменьшает шансы тараканов на поселение в квартире. Если бы мы получили нулевую корреляцию, то это означало бы, что никакой связи между наличием тараканов и ежедневной уборкой нет. Теперь можно перейти непосредственно к вычислению коэффициента ассоциации, которое ведется по следующей формуле:

$$r_a = \frac{ad - bc}{\sqrt{(a+b)(b+d)(c+d)(a+c)}}.$$

Подставим в эту формулу результаты наших наблюдений, приведенные в четырехпольной таблице.

$$r_a = \frac{5 \cdot 3 - 6 \cdot 6}{\sqrt{(5+6)(6+3)(6+3)(5+6)}} = -0{,}21.$$



Получилась отрицательная величина. Можно ли теперь утверждать, что ежедневное мытье полов приводит к уничтожению тараканов? Конечно же, пока нельзя! Мы же опросили не всех людей на земном шаре, а только двадцать друзей. Стало быть, нам необходимо определить достоверность вычисленного коэффициента. Это можно сделать с помощью все той же таблицы пороговых значений (табл. 3 в № 2 за 2008 год[2]). Для нашего случая $n = 20 - 2 = 18$, $P_{\text{дов}} = 95$ %. Теперь, используя $P_{\text{дов}}$ и n, находим в таблице III пороговую величину. Она равна $r_{\text{порог}} = 0{,}444$. После того как найдено значение $r_{\text{порог}}$, необходимо его сравнить с полученным эмпирическим коэффициентом, взятым без учета знака. Если эмпирическое значение больше порогового, то корреляция достоверна. Если меньше, то недостоверна. В первом случае мы можем утверждать, что связь между явлениями с вероятностью 95 % существует. Во втором — мы ничего утверждать не можем, можем лишь отметить, что достоверной связи не выявлено (что, впрочем, не означает отсутствие связи, она, может быть, выявится при увеличении объема выборки). Нетрудно заметить, что в нашем случае значение коэффициента ниже, чем пороговое. Стало быть, у нас нет оснований для вывода, что между ежедневной уборкой и наличием тараканов есть взаимосвязь.

Завершая разговор о корреляционном анализе, необходимо ответить на следующий вопрос. Для чего нам нужно знать силу и характер связи между явлениями?

Если мы действительно установили, что между явлениями есть взаимосвязь, то нам открывается огромное поле для всевозможных действий. Например, зная, что длина и вес у некоторого животного сильно взаимосвязаны, мы можем отказаться от трудоемкого процесса измерения и заниматься только взвешиваниями (в дальнейшем, пользуясь специальными методами, можно перевести вес в длину). Другой пример: мы показали, что коэффициент ассоциации между двумя видами организмов в некотором сообществе достоверен и имеет знак «–», что означает, что эти два организма избегают совместного поселения. Значит, между ними имеется некоторая биологическая связь (например, конкуренция, или они тяготеют к разным условиям среды). В этом случае корреляционный анализ позволил «нащупать» некоторое новое явление, которое далее можно изучать более внимательно. Наличие корреляции, также позволяет прогнозировать некоторые свойства. Если мы, скажем, установили, что между интенсивностью окраски плода и степенью его сладости существует высокая положительная корреляция, то процесс поиска наиболее сладких плодов заметно упрощается.

(Окончание следует)

2 В данной ситуации используется именно табл. 3, а не табл. 4.