УДК 575+311.3 https://doi.org/10.7124/FEEO.v28.1391

АТРАМЕНТОВА Л. А.
Харьковский национальный университет имени В. Н. Каразина,
Украина, 61022, г. Харьков, пл. Свободы, 4, e-mail: lubov.atramentova@gmail.com, (095) 66-26-736

СТАТИСТИЧЕСКИЙ АНАЛИЗ ЦИТОГЕНЕТИЧЕСКИХ ДАННЫХ

На примере данных, полученных в цитогенетическом исследовании, рассматриваются типичные ошибки, которые допускаются при выполнении статистического анализа. Широко распространенный, но некорректный статистический анализ неизбежно дает смещенный результат и повышает вероятность сделать неправильный научный вывод. Ошибки происходят из-за неучета дизайна исследования и структуры анализируемых данных. В статье показывается, как числовая несбалансированность комплекса данных приводит к смещению результата. На примере набора данных объясняется, как выполнить балансировку комплекса. Дается объяснение, в чем состоит преимущество представлять выборочные показатели с доверительными интервалами вместо статистических ошибок. Обращается внимание на необходимость при выборе статистического метода учитывать размер анализируемых долей. Показано, как один и тот же набор данных может быть поразному проанализирован в зависимости от цели исследования. Описывается алгоритм корректного статистического анализа, приводится форма табличного представления результатов.

Ключевые слова: структура данных, численно не сбалансированный комплекс, доверительный интервал.

В современной биологии и смежных с ней дисциплинах научный вывод обычно строится на основе статистических доказательств. Хотя в руководствах по статистике для биологов нет недостатка, их практическое применение часто оказывается проблематичным [1]. С началом компьютеризации у биологов появилась надежда на успешное решение проблемы. Исследователи, сняв с себя ответственность за результат статистического анализа, возложили ее на счетную технику. Чуда, однако, не произошло, но скорость, с которой стали получаться ошибочные результаты, возросла. Пришло осознание, что машина, делая сложные вычисления, не решает за исследователя научные задачи и что сбор научных фактов – это лишь первый этап в исследовании, за которым следует превращение фактов в данные и их анализ методами статистики.

Решение научной задачи предполагает ее формулировку в терминах статистики, учет схемы исследования, числа и объема групп, вида изучаемых признаков, структуры данных и многое другое, что до сих пор для многих биологов есть *terra incognita*. Причина этого – насильственно прерванная связь между отечественной биологией и мировыми разработками в области методологии научного исследования. От последствий этого разрыва отечественная наука не оправилась до сих пор [2], в чем можно убедиться, полистав журналы и авторефераты диссертаций по биологическим и смежным специальностям.

Цель данной статьи – на примере некорректного статистического анализа цитогенетических данных показать, как избежать распространенных ошибок.

Цитогенетические методы обычно применяют, чтобы выяснить, как влияют те или иные факторы на наследственный аппарат живых организмов. В экспериментах, преследующих цель выяснить действие фактора вообще, клетки служат экспериментальной моделью, а индивидуальные особенности их доноров в этом случае особой роли не играют и статистическим анализом не учитываются (схема 1). При изучении популяций растений, животных или человека важна именно индивидуальная реакция на экспериментальные или природные факторы, среди которых генетические и физиологические особенности доноров клеток (схема 2). Сказанное означает, что одни и те же данные в зависимости от цели и дизайна исследования требуют разного статистического подхода, хотя во всех случаях показателем служит удельный вес, или доля, клеток с искомыми специфичностями.

Ошибки статистического анализа часто обусловлены тем, что игнорируется структура данных, которые бывают простыми или состав-

ными. Простые данные получаются однократным измерением объекта. Данные, характеризующие объект несколькими измерениями, обозначают как составные. Цитогенетические данные в зависимости от схемы исследования могут быть проанализированы как простые (по схеме 1) или как составные (по схеме 2).

Рассмотрим анализ цитогенетических данных по схеме 1. В этом случае не учитываются индивидуальные особенности доноров, и данные представлены как простые, когда число наблюдений (количество клеток) равно числу единиц статистического учета. В табл. 1 приведен пример некорректного статистического анализа. Для нахождения среднего значения суммированы клетки всех объектов в группе, рассчитан процент клеток с искомой специфичностью – плюс-клеток. По формуле

$$s_{\%} = \sqrt{\frac{\%(100-\%)}{n-1}}$$

(n – количество клеток в группе) вычислены статистические ошибки и с помощью критерия Стьюдента определена статистическая значимость межгрупповой разницы. Расчеты, выполненные с использованием этого неподходящего к данному случаю алгоритма, привели к выводу, что изучаемый фактор повышает долю плюс-клеток в среднем до 16,81 % при 3,33 % в контроле, а пятикратная разница оказывается статистически значимой.

Описанный способ содержит серьезные методологические погрешности. Прежде всего, не учтено, что комплекс данных численно не сбалансирован – объекты представлены неодинаковым количеством клеток. В реальной научной работе получить сбалансированный комплекс данных, когда у всех объектов учтено одинаковое количество клеток, практически невозможно. Однако игнорировать это обстоятельство в статистическом анализе нельзя, иначе среднее значение будет неизбежно смещенным в сторону объекта с максимальным числом изученных клеток. В контрольной группе наибольшее влияние на результат оказывает объект № 3. Имея самый низкий процент плюс-клеток, он уменьшает среднее групповое значение. В опытном варианте наибольший вклад в среднее значение вносит объект № 6, из-за максимального процента плюс-клеток он смещает среднее значение в сторону увеличения. В результате разница между группами оказывается искусственно завышенной. Причина этого статистического артефакта – численная несбалансированность комплекса данных.

Чтобы избежать погрешности, комплекс данных балансируют, выравнивая вклад отдельных объектов в общий результат. Делают это по группам или по всему комплексу (табл. 2). Сумму клеток всех объектов делят на их число и каждому приписывают одно и то же среднее арифметическое значение n'. Затем вычисляют теоретическое количество плюс-клеток ($m = \frac{\% \times n'}{100}$), которые используют для расчетов корректных групповых средних. Как видим, эта процедура увеличила средний показатель в контрольной группе и уменьшила в опытной. Как следствие, разница между группами уменьшилась. Согласимся, изменения незначительны и на научный вывод не повлияли, но не нужно относиться к этому скептически – ведь известно, что культура в мелочах, а наука тем более.

Таблица 1. Некорректный способ статистического анализа

Группа	№ объекта	n	a	$\%_{\Sigma a} \pm s_{\%}$
Контрольная	1	300	14	
	2	350	11	
	3	400	10	
	Σ	1050	35	3,33±0,55
Опытная	4	150	23	
	5	200	25	
	6	340	68	
	Σ	690	116	16,81±1,42
Статистика		df = 2038; t = 8,87; $t_{0,001\,(\infty)}$ = 3,29; p<0,001		

Примечания: n – количество проанализированных клеток; a – количество плюс-клеток; Σ – сумма; $\%_a$ – процент плюс-клеток; $s_{\%}$ – статистическая ошибка процента; df – число степеней свободы; $t_{0,05(\infty)}$ и t – пороговое и фактическое значения критерия Стьюдента; p – уровень значимости.

Таблица 2. Балансировка комплекса данных

Исходные данные				Сбалансировано по группам			Сбалансировано по комплексу		
Группа	№	*n*	%	*n'*	*m*	%'	*n'*	*m*	%'
Контроль-ная	1 2 3	300 350 400	2,7 3,1 4,5	350 350 350	9,45 10,85 15,75	3,43	290 290 290	7,83 8,99 13,05	
	Σ	1050		1050	36,05		870	29,87	3,43
Опытная	4 5 6	150 200 340	15,3 12,5 20,0	230 230 230	35,19 28,75 46,00	15,93	290 290 290	44,37 36,25 58,00	
	Σ	690		690	109,94		870	138,62	15,93

Примечания: № – номер объекта; *n* – количество учтенных клеток; % – удельный вес плюс-клеток; Σ – сумма клеток; *n'* – сбалансированное количество учтенных клеток ($n' = \frac{\Sigma n}{k}$, *k* – число объектов); *m* – количество плюс-клеток среди сбалансированного общего количества ($m = \frac{\% \times n'}{100}$); *%'* – итоговый процент плюс-клеток.

Цитогенетические данные требуют особенно внимательного статистического анализа, так как из-за редкости цитогенетических событий может оказаться не выявленной реально существующая разница или, наоборот, может «обнаружиться» эффект, которого в действительности нет.

Другая ошибка анализа, представленного в табл. 1, состоит в том, что сравнение долей проведено без учета их размера. Цитогенетические показатели часто отражают редкие события и потому представлены малыми долями. Сравнение таких долей с использованием статистической ошибки дает смещенный результат. Корректный статистический анализ малых (≤ 20 %) и больших (≥ 80 %) долей выполняется переводом процентов в углы φ с последующим вычислением критерия Стьюдента $t = (\varphi_1 - \varphi_2)\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$ или Фишера $F = (\varphi_1 - \varphi_2)^2 \frac{n_1 n_2}{n_1 + n_2}$ (n_1 и n_1 – объемы сравниваемых групп). Этот прием универсальный, и его удобно использовать для сравнения долей любого размера.

В настоящее время представлять результаты статистического анализа в виде среднего показателя со статистической ошибкой является анахронизмом. Рационально для выборочных долей указывать доверительный интервал (табл. 3). Доверительные интервалы долей являются асимметричными, их трудно рассчитать вручную, но легко получить, воспользовавшись электронным калькулятором [3]. Преимущество доверительного интервала состоит в его наглядности. Читатель научной публикации легко может оценить значимость разницы между любыми выборочными долями, так как перекрывание доверительных интервалов свидетельствует о случайности выборочной разницы, а если доверительные интервалы не перекрываются, разница считается статистически значимой.

Обратимся к табл. 3. Доверительные интервалы у всех объектов в пределах групп перекрываются, указывая на их однородность, что является необходимым условием для объединения путем суммирования. В то же время доверительный интервал любого объекта контрольной группы не перекрывается с доверительными интервалами объектов опытной группы, что свидетельствует о статистически значимой разнице между отдельными объектами из разных групп.

В популяционном исследовании (схема 2) интерес представляют индивидуальные показатели, и в этом случае структура тех же данных выглядит иначе. Число единиц наблюдения равно числу клеток, а число статистического учета равно количеству объектов, в данном случае в каждой группе $n = 3$. В этой схеме проценты плюс-клеток анализируются не как доли, а как количественный непрерывный показатель, обозначаемый цитогенетиками как «количество случаев на 100 клеток» и требующий подхода, применяемого к количественным признакам: проверки на нормальность для решения вопроса о выборе между параметрической и непараметрической статистикой.

Таблица 3. Анализ индивидуальных показателей объектов

Группа	№	*n*	*a*	$\%_a$	95ДИ
Контрольная	1	300	14	4,67	2,28–7,05
	2	350	11	3,14	1,32–4,97
	3	400	10	2,50	0,97–4,03
Опытная	4	150	23	15,33	9,57–21,10
	5	200	25	12,50	7,92–17,09
	6	340	68	20.00	15,75–24,25

Примечания: № – номер объекта; *n* – общее количество проанализированных клеток; *a* – количество плюс-клеток; $\%_a$ – процент плюс-клеток; 95ДИ – 95-процентный доверительный интервал.

Таблица 4. Результаты анализа составных данных

Статистики	Контрольная группа	Опытная группа
n	3	3
lim	2,50–4,70	12,50–20,00
$\overline{x}$	3,43	15,93
s	1,13	5,33
$s_{\overline{x}}$	0,65	3,77
95ДИ	1,58–6,28	8,39–23,47
	$df = 4$; $t = 3{,}15$; $t_{0,05(4)} = 2{,}8$; $p<0{,}05$	

Примечания: *n* – количество объектов наблюдения; *lim* – минимальное и максимальное значения; $\overline{x}$ – среднее арифметическое значение; *s* – стандартное отклонение; $s_{\overline{x}}$ – статистическая ошибка среднего арифметического; *95ДИ* – 95%-процентный доверительный интервал; *df* – число степеней свободы; *t* и $t_{0,05(4)}$ фактическое и пороговое значения критерия Стьюдента; *p* – уровень значимости.

Анализ распределения данных в каждой группе указывает на отсутствие отклонения от нормального закона, что легитимизирует использование параметрической статистики. Как видим (табл. 4), средний показатель «количество случаев на 100 клеток» в контрольной и опытной группе точно совпадает с процентами, рассчитанными при анализе, когда комплекс данных рассматривался по схеме 1 как клеточная модель (табл. 2). Это указывает на корректность расчетов при использовании каждой из приведенных схем статистического анализа.

Еще одну некорректность статистического анализа можно обнаружить в публикациях, где встречаются абсолютные доли – 0 % и 100 %. Практически всегда авторы приводят абсолютные проценты без статистических ошибок, в то время как все другие проценты сопровождаются ими. Истоки этого понятны: для расчета статистической ошибки процента пользуются уже известной формулой, хотя она не пригодна для абсолютных долей, так как независимо от объема группы при расчетах всегда получается ноль. Ошибки для долей 0 и 100 % рассчитываются другим методом, хотя, как уже было сказано, принято приводить доверительные интервалы.

Приведённые способы статистического анализа применимы не только к цитогенетическим, но и к другим составным данным (количество потомков в семье, паразитов на животном, семян в плоде, зерен в колосе и многим другим).

References

1. Biometrica. Retrieved from: biometrica.tomsk.ru/index.htm. [in Russian].
2. They did not even imagine the science they took to abolish. Retrieved from: colta.ru/articles/society/18725. [in Russian].
3. Calculator of confidence limits for a sample proportion. Retrieved from: epitools.ausvet.com.au/ciproportion.

ATRAMENTOVA L. A.
V. N. Karazin Kharkiv National University,
Ukraine, 61022, Kharkiv, Svoboda sq., 4, e-mail: lubov.atramentova@gmail.com

STATISTICAL ANALYSIS OF CYTOGENETIC DATA

Using the data obtained in a cytogenetic study as an example, we consider the typical errors that are made when performing statistical analysis. Widespread but flawed statistical analysis inevitably produces biased results and increases the likelihood of incorrect scientific conclusions. Errors occur due to not taking into account the study design and the structure of the analyzed data. The article shows how the numerical imbalance of the data set leads to a bias in the result. Using a dataset as an example, it explains how to balance the complex. It shows the advantage of presenting sample indicators with confidence intervals instead of statistical errors. Attention is drawn to the need to take into account the size of the analyzed shares when choosing a statistical method. It shows how the same data set can be analyzed in different ways depending on the purpose of the study. The algorithm of correct statistical analysis and the form of the tabular presentation of the results are described.
Keywords: data structure, numerically unbalanced complex, confidence interval.

АТРАМЕНТОВА Л. О.
Харківський національний університет імені В. Н. Каразіна,
Україна, 61022, м. Харків, майдан Свободи, 4, e-mail: lubov.atramentova@gmail.com

СТАТИСТИЧНИЙ АНАЛІЗ ЦИТОГЕНЕТИЧНИХ ДАНИХ

На прикладі даних, отриманих у цитогенетичному дослідженні, розглянуто типові помилки, які допускають під час виконання статистичного аналізу. Широко розповсюджений, але некоректний статистичний аналіз неминуче призводить до зміщеного результату і підвищує ймовірність неправильного наукового висновку. Причинами помилок є неврахування дизайну дослідження і структури аналізованих даних. У статті з'ясовано, як числова незбалансованість комплексу даних призводить до зміщення результату. На прикладі набору даних пояснено, як виконати балансування комплексу. Показано, в чому полягає перевага представляти вибіркові показники з довірчими інтервалами замість статистичних помилок. Звернуто увагу на необхідність під час вибору статистичного методу враховувати розмір аналізованих часток. Показано, як один і той же набір даних залежно від мети дослідження може бути по-різному проаналізований. Описано алгоритм коректного статистичного аналізу, наведено форму табличного представлення результатів.
Ключові слова: структура даних, чисельно не збалансований комплекс, довірчий інтервал.