**DOI: 10.24412/2309-348X-2021-3-50-61**
**УДК 519.25:631.527**

# СТАТИСТИЧЕСКИЕ ОШИБКИ И КАК ИХ ИЗБЕГАЮТ, ИЛИ О КОРРЕКТНОМ АНАЛИЗЕ КОЛИЧЕСТВЕННЫХ ДАННЫХ В СЕЛЕКЦИИ

**А.А. СИНЮШИН**, кандидат биологических наук, ORCID ID: 0000-0003-4008-9460, E-mail: asinjushin@mail.ru

БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ МОСКОВСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА ИМЕНИ М.В. ЛОМОНОСОВА

*Все этапы селекционной работы предполагают получение и обработку количественных данных. Корректный статистический анализ позволяет делать адекватные выводы и сопоставлять результаты, полученные разными исследователями. При анализе публикаций в журнале «Зернобобовые и крупяные культуры» выявлены некоторые ошибки, наиболее часто встречающиеся при анализе количественных данных. Среди них использование параметрических методов (коэффициента корреляции Пирсона, t-критерия Стьюдента) без предварительной оценки характера распределения исследуемых признаков; неудачный выбор показателя центра распределения (среднее арифметическое вместо медианы); оперирование средними значениями без указания величины разброса; приведение результатов статистического анализа без оценки достоверности и некоторые другие. Отдельные процедуры анализа проиллюстрированы на примере собственных данных, полученных при описании коллекции гороха посевного. Предложен алгоритм, который может позволить снизить риск некорректного использования статистических приемов.*

***Ключевые слова***: анализ данных, корреляция, признак, распределение, статистика.

# STATISTICAL ERRORS AND HOW TO AVOID THEM, OR NOTES ON CORRECT ANALYSIS OF QUANTITATIVE DATA IN BREEDING

**A.A. Sinjushin**, ORCID ID: 0000-0003-4008-9460, e-mail: asinjushin@mail.ru
FACULTY OF BIOLOGY, LOMONOSOV MOSCOW STATE UNIVERSITY, MOSCOW, RUSSIA

***Abstract:*** *All stages of breeding process involve acquisition and processing of quantitative data. The correct statistical analysis provides a possibility to draw adequate conclusions and make results of different researches comparable. When analyzing papers published in 'Zernobobovye i Krupyanye Kul'tury' ('Legumes and Groat Crops'), we found several frequent errors in statistical analysis. Among them, one may list use of parametric methods (such as Pearson's correlation coefficient or Student's t-test) without preliminary evaluation of distribution of variables, ineffective choice of 'most typical' value (mean instead of median), use of means without illustrating the amount of variation, demonstrating the results of statistical analysis without estimates of their statistical significance etc. Some of analytical procedures were exemplified by own data from phenotyping of germplasm collection of a garden pea. An algorithm was also proposed which may reduce the risk of incorrect use of statistical methods.*

**Keywords**: correlation, data analysis, distribution, statistics, trait.

## Введение

Селекционная работа на всех этапах связана с получением и анализом количественных данных. Чтобы избежать необоснованных и субъективных оценок при работе с подобными данными, требуется прибегать к статистической обработке. Хотя основные статистические методы едины для различных областей знания, применительно к биологии в целом, генетике

и селекции существует определенная специфика, некогда отразившаяся в появлении термина «биометрия», – вероятно, все менее популярного в этом значении.

По мере самостоятельного освоения статистических методов любой исследователь сталкивается с двумя неочевидными обстоятельствами. Первое – набор статистических подходов во многом конвенционален, является предметом договоренности, а зачастую и традиции или вкуса. В отдельно взятой лаборатории для анализа корреляции могут пользоваться одним из нескольких возможных тестов – по сути, просто привычным. Даже в учебных пособиях по статистике встречаются суждения вида «принято считать» или «обычно используют». Это, однако, не означает, что все подходы взаимозаменяемы.

Второе – как ни парадоксально, на статистические методы тоже есть своя «мода». Так, традиционно во многих работах в качестве основного показателя значимости (достоверности) используют уровень значимости (*p*-value). Если выясняется, что различия двух массивов данных достоверны на 5%-ном уровне значимости ($p < 0,05$), это обычно считают признаком надёжности полученных результатов. Однако в течение нескольких последних лет использование *p*-value в качестве показателя «достоверности» всё чаще подвергается критике [1]. Таким образом, у статистических методов тоже есть своя динамика и эволюция.

При знакомстве с публикациями в отечественных научных и научно-производственных журналах, тематика которых включает генетику, селекцию и семеноводство, обращают на себя внимание некоторые наиболее распространенные ошибки в использовании статистических методов. Основная цель настоящей статьи – проанализировать эти ошибки и сформулировать рекомендации для корректного анализа.

Главным источником сведений о биометрии для многих поколений отечественных селекционеров остается классический труд Б.А. Доспехова «Методика полевого опыта» [2], выдержавший несколько переизданий. Спустя десятилетия после создания «Методики» сама процедура статистического анализа данных совершенно изменилась. Исчезла необходимость в кропотливых многостадийных расчетах вручную. Появились компьютерные программы, выполняющие сотни видов анализа (например, IBM SPSS, STATISTICA и многие другие). Несмотря на кажущуюся простоту машинной обработки результатов эксперимента, вопрос применимости тех или иных методов не стал менее существенным; возможно, напротив, оказался острее, потому что существенно расширился выбор из множества на первый взгляд сходных приемов.

В дополнение к «Методике» существует множество русскоязычных пособий по статистике – например, [3]. Существуют публикации, посвященные именно распространенным ошибкам в анализе данных. Источником вдохновения для настоящей работы послужила хорошо известная работа Тома Ланга, в которой рассмотрены статистические ошибки в медицинских публикациях [4]. К написанию данной статьи автора побудили многолетняя собственная работа, многократные эпизоды рецензирования рукописей и квалификационных работ, а также – необходимо признаться – опыт совершения некоторых из перечисленных ниже ошибок.

## Материал и методы исследований

В качестве модельных данных использованы собственные описания сортов и линий гороха (*Pisum sativum* L.) из коллекции кафедры генетики МГУ имени М.В.Ломоносова. Растения выращивали на опытном участке на территории Звенигородской биостанции МГУ (Московская область) в однорядном посеве. Подсчет числа семян в бобе производили при обмолоте. Массу семян определяли, взвешивая пять выборок по 10 семян на электронных весах Pioneer PA64 (Ohaus). Полученные значения умножали на 100 и таким образом получали традиционно используемый показатель «масса 1000 семян». Статистическую обработку данных выполняли с помощью программ Microsoft Excel 2016 (Microsoft Inc.) и STATISTICA 12 (StatSoft Inc.).

Для ознакомления с наиболее часто используемыми в селекционной работе методами были просмотрены статьи из журнала «Зернобобовые и крупяные культуры»,

опубликованные в 2020 г. Всего проанализированы 69 публикаций, в том или ином виде включающих анализ количественных данных.

**Результаты и их обсуждение**

Первичные результаты приведены в таблице 1.

Таблица 1

**Результаты описания сортов и линий гороха по признакам продуктивности**

| Сорт/линия | Семян в бобе[1], шт. | Масса 1000 семян[2], г | |
|---|---|---|---|
| | | 2015 г. | 2017 г. |
| Виола | 6; 6; 6; 6; 6; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7; 8; 8; 8; 8 | 154,2; 154,7; 162,2; 163,2; 172,1 | – |
| Батрак | 2; 2; 3; 3; 3; 4; 4; 4; 5; 5; 5; 5; 5; 5; 5; 5; 6; 6; 6 | 218,1; 220,0; 220,1; 229,9; 231,0 | 193,6; 199,0; 200,4; 202,8; 206,4 |
| Орел | 2; 2; 3; 3; 3; 4; 4; 4; 4; 4; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7 | 228,0; 252,2; 254,4; 255,5; 271,3 | 224,4; 234,4; 238,2; 239,0; 257,8 |
| Совершенство 65-3 | 4; 4; 4; 5; 5; 5; 5; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7; 7; 7 | 216,0; 218,0; 220,6; 227,2; 230,1 | 180,4; 205,4; 220,2; 227,1; 232,9 |
| Анвенд | 4; 5; 5; 5; 6; 6; 6; 7; 7; 7; 7; 7; 8; 8; 8; 8; 8; 8; 8; 9 | – | 234,6; 236,6; 239,4; 240,9; 263,2 |
| Малиновка | 1; 3; 4; 4; 4; 5; 5; 5; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7 | 180,0; 195,2; 195,6; 209,9; 211,8 | 150,6; 154,4; 158,7; 160,2; 161,3 |
| Roi des Gourmands | 3; 3; 3; 4; 4; 4; 5; 5; 5; 5; 5; 5; 5; 6; 6; 6; 6; 7; 7; 8 | – | 203,8; 231,6; 232,5; 234,9; 239,0 |
| Орлан | 4; 5; 5; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 8; 8; 9; 9; 9; 9 | 237,5; 265,5; 270,2; 275,3; 275,7 | 202,6; 203,3; 205,9; 206,0; 210,2 |
| Викинг | 2; 2; 4; 4; 4; 5; 5; 5; 5; 5; 5; 6; 6; 6; 6; 6; 6; 6; 6; 7 | 169,0; 185,3; 187,5; 187,8; 190,4 | 126,2; 134,4; 136,4; 143,9; 156,7 |
| Мультик | 4; 4; 4; 5; 5; 5; 6; 7; 7; 7; 7; 7; 7; 7; 7; 7; 7; 8; 9; 9 | 155,4; 156,0; 156,2; 160,7; 167,3 | 124,8; 124,9; 126,1; 127,8; 128,5 |
| SGE | 2; 4; 5; 5; 5; 5; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8 | 150,6; 158,6; 163,3; 183,2; 186,8 | 156,7; 157,3; 158,5; 180,4; 189,0 |

*[1]Результаты описания 20 бобов для каждого образца. [2]Результаты 5 взвешиваний для каждого образца, умноженные на 100. Прочерк – данные отсутствуют.*

Ниже мы сформулируем несколько наиболее часто встречающихся ошибок и укажем на способы их обойти.

**Ошибка 1. Использование статистических методов без их описания.** По требованиям журналов использованные методы (в том числе описание статистической обработки данных) требуется характеризовать со степенью детализации, достаточной для воспроизведения. Однако в значительном числе просмотренных публикаций было указано лишь, что статистическая обработка данных выполнена по Б.А. Доспехову, что, безусловно, не информативно: в книге [2] приводится описание десятков статистических процедур. В трех статьях нами найдено утверждение, что проведен дисперсионный анализ в соответствии с руководством [2], но в тексте никаких атрибутов или результатов дисперсионного анализа нет. Некоторые авторы ограничились ссылками на «общепринятые» или «стандартные» методики без указания источников.

Для адекватного восприятия результатов статистического анализа обязательно приводить сведения о размере выборки, числе повторностей, использованных статистических критериях и тестах (например, так: «взаимосвязь между признаками оценивали с помощью коэффициента корреляции Спирмена»). Эта информация гораздо важнее, чем отсылка к какому-либо справочному изданию и даже указание программы, в которой были произведены расчеты.

**Ошибка 2. Включение в анализ выпадающих точек.** В массиве данных могут содержаться значения (например, полученные ошибочно или характеризующие нетипичные, аномальные растения), которые сильно искажают результат, если их своевременно не исключить. В пособии Б.А. Доспехова для решения этой задачи предлагается критерий τ (тау). Существует упрощенный подход, в соответствии с которым выпадающими (выбросами, outliers) считаются значения, не попадающие в интервал [Q1 – 1,5·IQR; Q3 + 1,5·IQR], где Q1, Q3 и IQR – соответственно первый квартиль, третий квартиль и интерквартильный размах.

В современных реалиях наиболее быстрый способ найти выпадающие значения и оценить характер распределения – построить график вида «усатый ящик» (box and whiskers plot). Такие графики для признака числа семян в бобе у четырех сортов были построены с использованием пакета STATISTICA (рис. 1). У сорта Батрак верхний квартиль совпадает с медианой, у сорта Совершенство 65-3 – с максимальным значением, у Roi des Gourmands распределение близко к симметричному, а значение «2» для сорта Викинг является выпадающим – его мы исключили в дальнейшем анализе. Аналогично, выпадающим оказалось значение «4» для числа семян у сорта Орлан (график не представлен), оно тоже подлежит исключению.

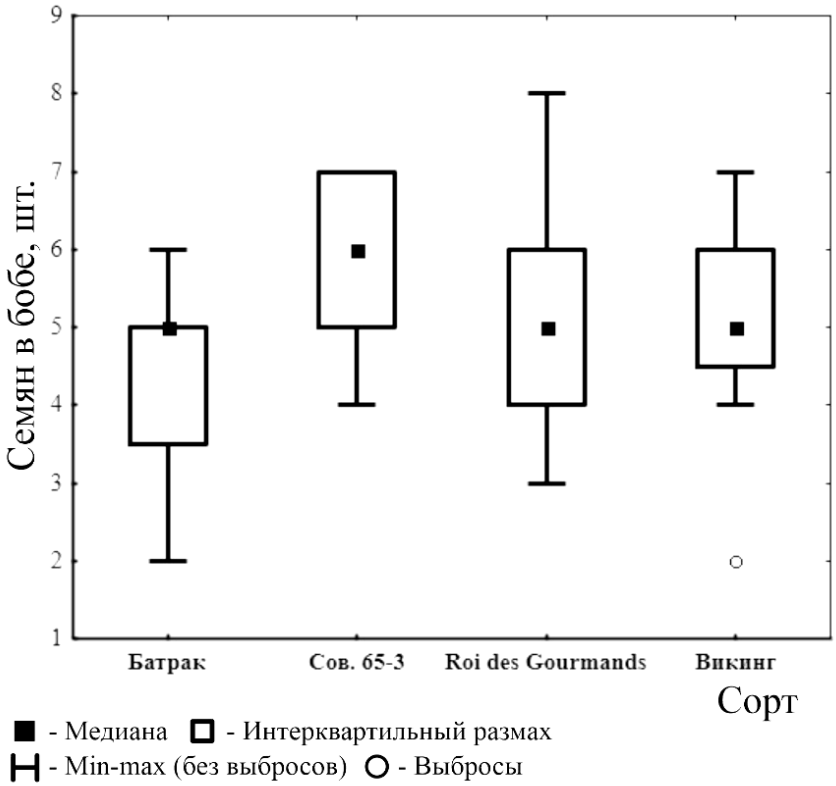


*Рис. 1. Распределение числа семян в бобе у четырех сортов*

Помимо выбросов, при анализе могут быть обнаружены пограничные значения (extremes), находящиеся на «краях» распределения. Вопрос о включении таких значений в выборку для дальнейшего анализа не имеет общепринятого ответа.

Практика исключения выбросов в некоторых случаях представляется небезупречной с биологической точкой зрения. Так, значение «2» (формально выпадающее) для признака числа семян в бобе встретилось у двух растений из 20 в сорте Викинг (т.е. в 10% случаев; табл. 1), и его сложно считать «аномальным» или «нетипичным». Исключение таких случаев чревато искусственным сужением реально существующего разнообразия. Альтернативой

может стать более обоснованный, чем подсчет среднего арифметического, выбор центра распределения, о чем пойдет речь дальше.

**Ошибка 3. Выбор центра распределения без учета характера распределения.** Коль скоро значения признака распределены в соответствии с некоторой закономерностью, требуется максимально эффективно описать «наиболее типичное» значение и то, насколько сильно отдельные элементы отклоняются от него.

Чаще всего на практике используют среднее арифметическое ($\bar{x}$). Очевидно, что $\bar{x}$ «типично» только для симметричных распределений – например, для нормального. Однако, даже если в целом (в генеральной совокупности или достаточно большой выборке – обычно не меньше 50 значений) значение признака распределено нормально, в малых выборках отклонения могут быть существенными (рис. 1, 2).

Уместно заметить, что среднее – это число, точка на координатной оси, но никак не диапазон. Формулировки вида «средняя 5,5-6,5 см» некорректны.

Одна из наиболее распространенных альтернатив среднему арифметическому – медиана ($M$): такое число, что половина всех значений в выборке не меньше $M$, а другая половина – не больше. Именно медиана во многих случаях надёжно отражает «самое типичное» значение в выборке, хотя это не так для некоторых необычных распределений – например, симметричного с двумя пиками.

Еще один показатель центра распределения – мода ($Mo$): наиболее часто встречающееся в выборке значение. Если признак распределен непрерывно (так обстоит дело с длиной, высотой, содержанием белка в зерне, массой и т.д.), мода может не иметь смысла, поскольку теоретически каждый образец может характеризоваться уникальным значением признака. Однако для признаков с непрерывным распределением часто используют разбиение на классы (например, карликовые, полукарликовые, среднерослые и т.д.), и в этом случае мода информативна.

Для дискретных величин (количество семян, число дней до цветения и пр.) мода – очень наглядный показатель. В отличие от $\bar{x}$ и $M$, мода пригодна и для описания качественных (номинальных) переменных: например, в группе из 70 зерновых сортов, 20 овощных и 10 фуражных модой будет значение «зерновой сорт».

В табл. 2 приведены показатели значения среднего, медианы и моды для количества семян в бобе у проанализированных сортов. Из приведенных чисел видно, что мода и медиана гораздо более устойчивы к выбросам. Даже если искусственно включить в анализируемую выборку заведомо выпадающее значение (например, добавить число 1000 к количествам семян в бобе из табл. 1), это сильно отклонит среднее значение, не изменит моду и, возможно, незначительно (в нашем случае не более чем на 0,5) сместит медиану. Поэтому использование медианы (а не среднего арифметического) как показателя центра распределения – удобная альтернатива дополнительному анализу на наличие выбросов.

Как видно из таблицы 2, выборка может иметь мультимодальное распределение, т.е. иметь более чем одну моду – так обстоит дело для линии SGE. Это еще одно ограничение на использование моды в расчетах. Однако именно мода для дискретно распределенных величин имеет интуитивно наибольший биологический смысл. Она принимает только те значения, которые есть в выборке и, например, не может стать дробной, если признак имеет только целочисленные значения.

Если значения признака имеют нормальное распределение, в генеральной совокупности среднее арифметическое и медиана совпадают (и близки в выборке). Однако значительная часть признаков (или выборок, особенно малых) распределены иначе. Существуют строгие критерии, с помощью которых можно оценить, насколько нормально распределен признак (критерии Шапиро-Уилка, Колмогорова-Смирнова и др.). На практике гораздо более простая альтернатива – использовать в качестве показателя центра распределения медиану, а не среднее арифметическое.

Здесь необходимо заметить, что в селекционной работе многие процедуры регламентируются государственными стандартами (ГОСТ) или иного рода указаниями

(например, [5]), которые предполагают использование только среднего арифметического. Прочие показатели центра распределения (например, табл. 2) могут быть наглядным дополнением при обсуждении полученных результатов.

Таблица 2

**Показатели центра распределения и рассеяния для признака числа семян в бобе**

| Сорт/линия | Семян в бобе, шт. | | | | | |
|---|---|---|---|---|---|---|
| | $\bar{x}$ ± SD (CV) | | *M* (Q1; Q3) | | *Mo* (min-max) | |
| Виола | 6,8 ± 0,8 (11,8%) | | 7,0 (6,0; 7,0) | | 6 (6-8) | |
| Батрак | 4,4 ± 1,2 (27,3%) | | 5,0 (3,5; 5,0) | | 5 (2-6) | |
| Орел | 4,8 ± 1,7 (35,4%) | | 4,5 (3,5; 6,0) | | 4 (2-7) | |
| Сов. 65-3 | 5,9 ± 1,2 (20,3%) | | 6,0 (5,0; 7,0) | | 7 (4-7) | |
| Анвенд | 6,9 ± 1,3 (18,8%) | | 7,0 (6,0; 8,0) | | 8 (4-9) | |
| Малиновка | 5,4 ± 1,6 (29,6%) | | 5,5 (4,5; 7,0) | | 7 (1-7) | |
| Roi des G. | 5,1 ± 1,4 (27,5%) | | 5,0 (4,0; 6,0) | | 5 (3-8) | |
| Орлан | 7,2 ± 1,4 (19,4%)[1] | 7,4 ± 1,3 (17,6%)[2] | 7,0 (6,5; 8,0)[1] | 7,0 (7,0; 8,0)[2] | 7 (4-9)[1] | 7 (5-9)[2] |
| Викинг | 5,1 ± 1,3 (25,5%)[1] | 5,4 ± 0,8 (14,8%)[2] | 5,0 (4,5; 6,0)[1] | 5,5 (5,0; 6,0)[2] | 6 (2-7)[1] | 6 (4-7)[2] |
| Мультик | 6,5 ± 1,5 (23,1%) | | 7,0 (5,0; 7,0) | | 7 (4-9) | |
| SGE | 5,9 ± 1,4 (23,7%) | | 6,0 (5,0; 7,0) | | 5 и 7 (2-8) | |

*[1] Включая выбросы. [2] Не включая выбросы.*

**Ошибка 4. Приведение только средних значений без характеристики рассеяния.** В очень многих (39 из 69 просмотренных) публикациях в обобщающих таблицах для признаков приведены только средние значения – например, за несколько лет полевых опытов. Такое представление не отражает, насколько стабильно воспроизводится это среднее значение, а также насколько существенен разброс индивидуальных значений, хотя эта информация чрезвычайно важна при оценке результатов селекционной работы. М.А. Вишнякова и др. [5: с. 87] замечают, что «прямое сравнение средних арифметических признаков разных генотипов не всегда правомерно, так как условия их произрастания сильно варьируют», и с этим нельзя не согласиться.

Необходимо тем или иным образом характеризовать рассеяние значений признака. Возможны несколько вариантов (табл. 2).

1. Стандартное отклонение (*s*, standard deviation, SD). Эта величина совместима со средним арифметическим и для нормального (или близкого к нему, см. ниже) распределения отражает, насколько сильно отдельные значения в выборке отклоняются от среднего. Во многих случаях в таблицах значения признака приводят как $\bar{x}$ ± SD.

2. Коэффициент вариации (coefficient of variation, CV), т.е. отношение стандартного отклонения к среднему значению, обычно выражаемое в процентах. В отличие от SD, это безразмерная величина, и ее часто используют, чтобы минимизировать «эффект шкалы». В самом деле, для длины стебля SD = 2,5 см – очень небольшая величина, а для длины боба гораздо более существенная. Высокий CV может указывать на неэффективность селекции по данному признаку из-за очень широкой нормы реакции.

3. Интерквартильный размах (interquartile range, IQR), приводимый как разность третьего (Q3) и первого (Q1) квартилей или – что более информативно – просто как их значения. Этот параметр хорошо характеризует асимметрично распределенные величины и совместим с медианой. Приведение данных в виде $M \pm 0{,}5 \cdot \mathrm{IQR}$ некорректно, поскольку квартили могут быть на разных расстояниях от медианы и даже совпадать с ней (рис. 1).

4. 95%-ный доверительный интервал (confidence interval, CI). Данный показатель используется редко (одна публикация из просмотренных нами), хотя в целом хорошо

характеризует выборки из популяции, в которой значение признака распределено нормально. В ином случае не обоснован выбор t-статистики для оценки доверительного интервала.

5. Наибольшее и наименьшее значение признака. Этот способ используют чаще прочих (пять публикаций из 69), но существует риск, что именно здесь будут приведены нетипичные, уклоняющиеся значения (см. выше).

В качестве визуального представления можно использовать график вида «усатый ящик» (рис. 1), а также добавлять «усы» к гистограммам или диаграммам рассеяния.

Стандартную ошибку среднего (standard error of mean, SEM, SE) некорректно использовать как показатель рассеяния в выборке, хотя SEM иногда приводят на тех же правах, что и стандартное отклонение [4, 6]. SEM всегда меньше SD и потому выглядит несколько более привлекательно, но имеет совершенно иной смысл и отражает то, насколько точно среднее в выборке соответствует среднему в генеральной совокупности.

**Ошибка 5. Использование t-критерия Стьюдента для сравнения малых или асимметрично распределенных выборок.** Одна из наиболее часто встречающихся задач – сравнение выборок с целью установить, достоверно ли они различаются. Некорректно сравнение просто по средним значениям, хотя во многих публикациях встречаются формулировки вида «внесение удобрений привело к повышению урожайности по сравнению с контролем на 5%» без оценки достоверности этих различий.

Для сравнения существуют многочисленные критерии, из которых наиболее популярен t-критерий Стьюдента. Важно, однако, что выборочные средние имеют распределение Стьюдента (оно симметрично), если в генеральной совокупности признак распределен нормально. Иными словами, использовать t-критерий можно, если сравниваемые данные в масштабах генеральной совокупности (например, в популяции) имеют нормальное распределение, что далеко не всегда встречается на деле.

Построим гистограммы распределения для массы 1000 семян урожаев 2015 и 2017 гг. В обоих случаях признак имеет несимметричное, явно отличное от нормального распределение (рис. 2). Если применить для сравнения этих двух выборок критерий Стьюдента, получим, что различия достоверны ($p < 0{,}01$). Более корректным будет использовать непараметрический критерий. Для связанных (зависимых) переменных (именно таковы массы семян одних и тех же сортов в разные годы) пригоден Т-критерий Уилкоксона, который также показывает, что различия достоверны, хотя и на более низком уровне значимости ($p < 0{,}05$).

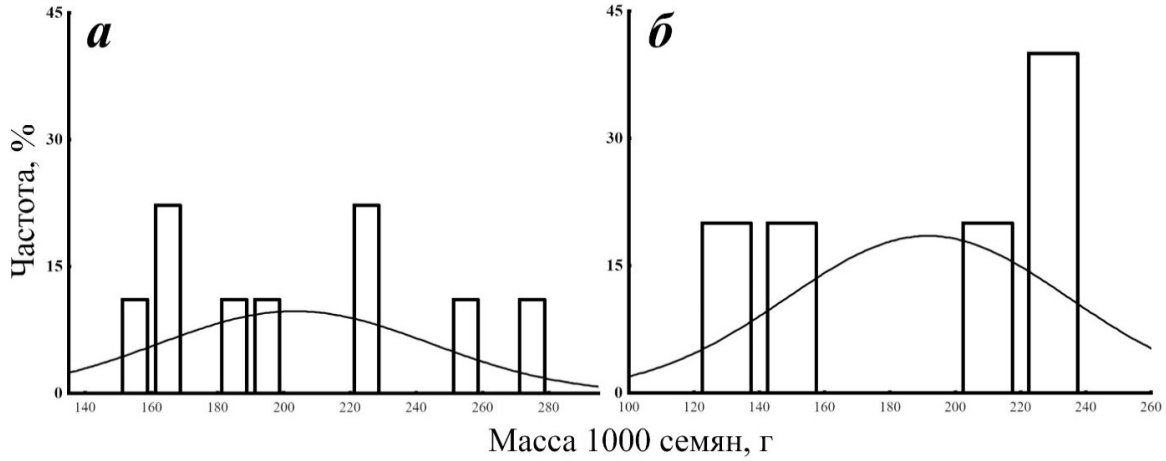


*Рис. 2. Частотные гистограммы для признака «масса 1000 семян» у группы сортов (проанализированы медианные значения для каждого сорта) в 2015 (а) и 2017 (б) гг. Для сравнения показаны графики функции плотности вероятности нормального распределения, построенные для тех же средних значений и стандартных отклонений*

Вопрос о применимости тех или иных статистических методов к признакам, имеющим различное распределение, во многом решается на уровне традиции. Обычно прибегают к следующим допущениям.

1. Если выборка достаточно большая (50 и более значений), применяют t-критерий.
2. Если признак имеет дискретные значения, но в широком диапазоне (например, со значениями от 10 до 30), его распределение может быть близким к нормальному.
3. Если о признаке известно (например, по предыдущим исследованиям на больших выборках), что в популяции он распределен нормально, t-критерий применяют и к выборкам небольшого объема из этой популяции.

К последнему аргументу следует прибегать с осторожностью, поскольку генеральная совокупность при исследовании одного и того же признака может быть различной в зависимости от задачи. Например, длина стебля у одного сорта, скорее всего, распределена нормально. Однако в большой коллекции сортов тот же самый признак может иметь иное распределение – например, бимодальное, соответствующее длинно- и короткостебельным сортам.

**Ошибка 6. Использование одних и тех же критериев для сравнения двух и нескольких выборок.** Сравнивая несколько выборок друг с другом на предмет достоверности различий, категорически неверно использовать тот же критерий, что и для попарного сравнения, – например, критерий Манна-Уитни или t-критерий Стьюдента. Существуют специальные критерии для множественного сравнения. Для одновременного сравнения трех и более небольших выборок используют критерий Краскела-Уоллиса, который проверяет гипотезу о том, что различий между группами в принципе нет. Для попарного сравнения «всех со всеми» можно использовать критерий Данна.

Определим, достоверно ли различаются по числу семян в бобе сорта Виола, Батрак, Орел и Малиновка (табл. 3). Выбранные критерии не являются взаимозаменяемыми: критерий Манна-Уитни при попарном сравнении показывает достоверные различия в большем числе случаев, чем критерий Данна, тем самым создавая ошибочную картину.

Таблица 3

**Уровни значимости (*p*-values) при попарном сравнении сортов гороха по признаку числа семян в бобе**

| | Виола | Батрак | Орел | Малиновка |
|---|---|---|---|---|
| Виола | – | 0,000** | 0,000** | 0,004** |
| Батрак | 0,000** | – | 0,525 | 0,033* |
| Орел | 0,000** | 1,000 | – | 0,223 |
| Малиновка | 0,033* | 0,218 | 1,000 | – |

*В верхней правой части таблицы приведены результаты попарного сравнения критерием Манна-Уитни, в нижней левой – множественного сравнения критерием Данна. Звездочками отмечены достоверные различия (* $p < 0,05$; ** $p < 0,01$).*

**Ошибка 7. Использование коэффициента корреляции Пирсона для анализа малых выборок или нелинейной зависимости.** В 11 публикациях авторы постулировали существование корреляции между анализируемыми величинами, не уточняя, какой именно коэффициент корреляции был использован. Указание на использование программы Microsoft Excel позволяет предположить, что речь о функции КОРРЕЛ, которая вычисляет значение коэффициента корреляции Пирсона.

Вычисление этого коэффициента ($r$) корректно только для нормально распределенных величин с уже перечисленными выше оговорками. Он также характеризует лишь линейную зависимость. На практике часто встречаются параметры, связанные нелинейно, – например, график зависимости эффекта от дозы часто представляет собой S-образную кривую насыщения. Расчет коэффициента корреляции Пирсона некорректен, если заранее неизвестны законы распределения двух переменных и характер зависимости между ними. Альтернативой являются непараметрические методы, которые применимы к выборкам

малого объема, не требуют информации о характере распределения величин, способны обнаруживать любые монотонные зависимости и менее чувствительны к выбросам. Чаще всего используются коэффициенты корреляции Спирмена ($\rho$) и Кендалла ($\tau$).

Например, между медианными значениями массы 1000 семян, полученных в 2015 и 2017 гг. (т.е. для восьми пар значений, распределенных не нормально: рис. 2), коэффициент корреляции Пирсона (применяемый здесь некорректно!) $r$ = 0,853; коэффициент Спирмена $\rho$ = 0,905; коэффициент Кендалла $\tau$ = 0,786. Таким образом, использование непараметрических методов может дать даже более «победоносный» результат и, безусловно, гораздо более обоснованно в данном случае. Корреляция между одними и теми же данными, полученными в разные годы (автокорреляция), вероятно, может служить приблизительной оценкой воспроизводимости этих данных.

Во многих случаях характер зависимости (линейный или нелинейный) понятен по точечной диаграмме рассеяния (например, рис. 3). На такой диаграмме можно увидеть и резко уклоняющиеся от общей тенденции точки, которые могут существенно влиять на интерпретацию. Показателен в этом смысле знаменитый «квартет Энскомба» [7] – четыре набора данных, которые имеют ряд одинаковых показателей (средние, дисперсию, коэффициенты корреляции и детерминации), но очень различаются по характеру зависимости. Поэтому во многих случаях наглядно представить данные в виде графика рассеяния; ни в одной из просмотренных нами работ подобных графиков не было.

Независимо от выбранного метода оценки корреляции, необходимо помнить, что она отражает лишь статистическую – но необязательно биологическую и/или причинно-следственную связь между параметрами.

**Ошибка 8. Использование коэффициента детерминации ($R^2$) вместо коэффициента корреляции.** При отображении зависимости между двумя переменными на диаграмме рассеяния иногда добавляют линию регрессии (тренда). Эта линия представляет собой график функции, с той или иной степенью точности аппроксимирующей экспериментальные данные. Точность аппроксимации выражают в виде коэффициента детерминации ($R^2$), который в случае линейной зависимости равен квадрату коэффициента корреляции Пирсона ($r$) между анализируемыми переменными.

Коэффициент детерминации отражает только то, насколько хорошо экспериментальные данные соответствуют модели (описываются предложенным уравнением регрессии), но – в общем виде – не силу статистической взаимосвязи между двумя наборами чисел. Рассмотрим пример со случайными данными, нелинейно зависимыми друг от друга (рис. 3).

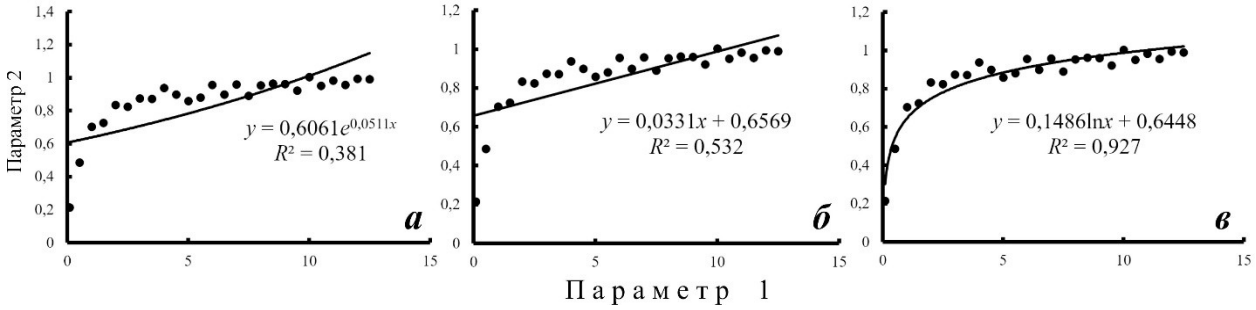


*Рис. 3. Аппроксимация одних и тех же нелинейно зависимых данных экспоненциальной (а), линейной (б) и логарифмической (в) моделями*

Коэффициент корреляции Пирсона (применяемый здесь некорректно!) для этих данных составляет 0,729, коэффициент корреляции Спирмена (непараметрический, т.е. подходящий к этой ситуации) равен 0,902. Однако $R^2$ существенно различается для разных моделей регрессии. Экспоненциальная зависимость (рис. 3, а) описывает эти данные плохо ($R^2$ = 0,381, т.е. лишь 38,1% дисперсии зависимой переменной объясняется моделью), немногим более 50% дисперсии могут быть объяснены линейной моделью (рис. 3, б), а

логарифмическая модель аппроксимирует данные с наибольшим успехом (рис. 3, в). Однако речь идет об одних и тех же исходных данных, т.е. $R^2$ никак не описывает силу статистической взаимосвязи между переменными и не должен быть использован с этой целью.

**Ошибка 9. Приведение результатов статистической обработки без оценки достоверности.** Недостаточно указать, что значение признака в одной группе больше, чем в другой, или что два параметра коррелируют. Использование статистических методов предполагает оценку того, насколько достоверны различие или связь.

Как отмечают М.А. Вишнякова и др. [5: с. 87], «выявление лучших сортов… чаще всего проводят самым простым и распространенным методом – сравнением средних арифметических величин… полученных для каждого варианта опыта, без установления достоверности результатов». Сложно признать сложившуюся практику безупречной.

В большинстве отечественных работ (42 из 69 просмотренных публикаций) в качестве оценки достоверности используется наименьшая существенная разность (НСР, least significant difference, LSD). Этот параметр указывает на границу возможных случайных отклонений в эксперименте, за пределами которой различия средних считаются значимыми на соответствующем уровне. На практике обычно используют 1%-ный и 5%-ный уровни значимости ($НСР_{01}$ и $НСР_{05}$). Важно, однако, что расчет НСР обращается к t-критерию Стьюдента и, следовательно, информативен не для любых распределений (см. выше).

При анализе данных с помощью специальных программ обычно автоматически рассчитывается уровень значимости, выраженный через *p*-value. В большинстве естественнонаучных исследований обращают внимание на два порога: $p < 0{,}05$ и $p < 0{,}01$. В последние годы обострилась дискуссия о том, насколько оправдано использование *p*-value в качестве показателя достоверности [1]. Указывают, что этот показатель очень чувствителен к размеру выборки и зачастую плохо воспроизводим. Можно констатировать некоторую смену традиции, но приведение *p*-value является гораздо более удачным решением, чем отсутствие какой-либо оценки достоверности.

**Заключение**

Обобщим все вышеизложенное в виде алгоритма, который, возможно, позволит обработать и представить данные более корректно и эффективно. При работе с количественными признаками рекомендуется выполнить следующие действия.

1. Выполнить проверку на выпадающие точки.

2. Определить характер распределения исследуемых признаков. Если выборки небольшие (меньше 20) или распределение резко отличается от нормального, для дальнейшего анализа необходимо использовать непараметрические методы. Наличие выпадающих точек или соответствие распределения нормальному можно достаточно примерно оценить, вычислив для выборки среднее значение и медиану: у симметричных распределений без выбросов эти показатели близки.

3. При анализе корреляции удобно построить точечную диаграмму рассеяния, чтобы оценить линейность зависимости между переменными. В случае отчетливо нелинейной зависимости (равно как и для малых и асимметрично распределенных выборок) корректнее использовать непараметрические коэффициенты корреляции.

4. Приводя в тексте, таблицах или на графиках средние показатели, необходимо тем или иным образом указывать величину разброса.

5. При любых статистических выкладках обязательно отображать, насколько значимы показатели (например, с помощью *p*-value).

В таблице 4 приведены примеры статистических методов с указанием областей их применения.

Данный небольшой обзор ни в коем случае не преследовал цель обучения статистике – это заняло бы гораздо больший объем и в целом избыточно в мире, где существует множество специализированных учебных изданий. Здесь рассмотрены лишь некоторые ошибки, которые периодически встречаются в отечественных работах по селекции.

Таблица 4

**Методы, пригодные для решения некоторых статистических задач**

| Задача | | Непараметрические методы | | Параметрические методы | |
|---|---|---|---|---|---|
| Поиск корреляции | | коэффициенты корреляции Спирмена, Кендалла | | коэффициент корреляции Пирсона | |
| Оценка достоверности различий | | Две переменные | Более двух переменных | Две переменные | Более двух переменных |
| | Связанные значения | критерий Уилкоксона | критерий Фридмана | t-критерий Стьюдента для связанных значений | парный t-критерий для множественных сравнений |
| | Несвязанные значения | критерий Манна-Уитни | критерии Краскела-Уоллеса, Данна | парный t-критерий Стьюдента | дисперсионный анализ (ANOVA), критерий Тьюки |

Грамотный выбор процедуры анализа необходим в селекционной работе. Тщательно подобранные и детально описанные статистические методы позволят не только получить обоснованные и адекватные выводы, но и сделать результаты различных исследований сопоставимыми между собой – в том числе и на общемировом уровне. Уверенное владение статистикой способствует международному восприятию и признанию отечественных журналов и опубликованных в них работ гораздо более эффективно, чем требования к авторам дублировать список литературы и заглавия таблиц на английском языке. Хочется надеяться, что настоящая работа будет полезна исследователям и практикам в области селекции в их повседневной деятельности.

**Благодарности**

Автор сердечно благодарит к.б.н. С.Н. Лысенкова за многолетние беседы о статистике, критический просмотр данной рукописи и неоценимые комментарии, а также рецензента, д.с.-х.н. А.Н. Зеленова за конструктивные замечания.

**Литература**

1. Nuzzo R. Scientific method: statistical errors //Nature. – 2014. – V. 506. – № 7487. – P. 150-152. DOI: 10.1038/506150a

2. Доспехов Б.А. Методика полевого опыта (с основами статистической обработки результатов исследований). 5-е изд. – М.: Агропромиздат. – 1985. – 351 с.

3. Гланц С. Медико-биологическая статистика. – М.: Практика. – 1998. – 459 с.

4. Ланг Т. Двадцать ошибок статистического анализа, которые вы сами можете обнаружить в биомедицинских статьях // Междунар. журн. мед. практики. – 2005. – №1. – С. 21-31.

5. Вишнякова М.А., Сеферова И.В., Буравцева Т.В. и др. Коллекция мировых генетических ресурсов зерновых бобовых ВИР: пополнение, сохранение и изучение: методические указания. 2-е изд. – СПб.: ВИР. – 2018. – 143 с.

6. Altman D.G., Bland J.M. Standard deviations and standard errors // BMJ. – 2005. – V. 331. – P. 903. DOI: 10.1136/bmj.331.7521.903

7. Anscombe F.J. Graphs in statistical analysis // Am. Stat. – 1973. – V. 27. – P. 17-21. DOI: 10.2307/2682899

**References**

1. Nuzzo R. Scientific method: statistical errors. Nature, 2014, vol. 506, no. 7487. – P. 150-152. DOI: 10.1038/506150a

2. Dospekhov B.A. Metodika polevogo opyta (s osnovami statisticheskoi obrabotki rezul'tatov isslefovanii) [Methods of field experiments]. 5th ed. Moscow: Agropromizdat, 1985, 351 p. (In Russian).
3. Glantz S. Mediko-biologicheskaya statistika [Primer of biostatistics]. Moscow: Praktika, 1998, 459 p. (In Russian).
4. Lang T. Twenty statistical errors even *YOU* can find in biomedical research articles. Croat. Med. J., 2004, vol. 45, no. 4, pp. 361-370.
5. Vishnyakova M.A., Seferova I.V., Buravtseva T.V. et al. Kollektsiya mirovykh geneticheskikh resursov zernovykh bobovykh VIR: popolneniye, sokhraneniye i izucheniye: metodicheskiye ukazaniya [VIR global collection of grain legume crop genetic resources: replenishment, conservation and studying]. Saint-Petersburg, VIR, 2018, 143 p. (In Russian).
6. Altman D.G., Bland J.M. Standard deviations and standard errors. BMJ, 2005, vol. 331, pp. 903. DOI: 10.1136/bmj.331.7521.903
7. Anscombe F.J. Graphs in statistical analysis. Am. Stat., 1973, vol. 27, pp. 17-21. DOI: 10.2307/2682899