STUDIA UNIVERSITATIS MOLDAVIAE, 2015, nr.6(86)
Seria "Ştiinţe reale şi ale naturii" ISSN 1814-3237 ISSN online 1857-498X p.119-124

ДИСПЕРСИОННЫЙ АНАЛИЗ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ В АГРОНОМИЧЕСКОМ ПОЛЕВОМ ЭКСПЕРИМЕНТЕ И ЕГО СВЯЗЬ С ЗАКОНОМ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Андрей БАБИЦКИЙ

Славянский университет, г. Кишинев

При сравнении дисперсионного анализа с законом нормального распределения разработан новый математически облегченный метод дисперсионного анализа для биологов, позволяющий в полевых агрономических опытах выделять влияние регулируемого воздействия на растения на фоне нерегулируемых шумовых воздействий факторов внешней среды. Описан смысл дисперсионного анализа.

Ключевые слова: *дисперсионный анализ, разложение дисперсии, центрирование дат, нормальное распределение.*

ANALYSIS OF VARIANCE OF QUANTITATIVE TRAITS IN AGRONOMIC FIELD EXPERIMENTS AND ITS CONECTION WITH THE LAW OF NORMAL DISTRIBUTION

A new, mathematically easier for biologists, the analysis of variance procedure as an alternative approach in the study of the degree of influence of various controlled factors on the plant in the agronomic and biological research under uncontrolled field environments has been developed.. The degree of influence of the studied factors on the plant is given in percentage terms. Meaning analysis of variance has been described.

Keywords: *analysis of variance, variance decomposition, centering dates, normal distribution.*

Для изучения ответной реакции растения на влияние на него определенного по градациям фактора, важно отделить влияние других факторов внешней среды, которые в данный момент также влияют на растение, но являются факторами шума, который может перекрыть влияние изучаемого фактора. Для этих целей в настоящее время применяют климатические камеры или фитотрон, где контролируется температура, фотопериод, освещенность и другие параметры внешней среды.

В полевых агрономических экспериментах невозможно контролировать все условия внешней среды, в которых находится растение. Четко изучить влияние нужного фактора на растения подчас весьма трудно, ибо не удается выделить это влияние на фоне влияния других факторов внешней среды. Многие факторы внешней среды вызывают не контролируемую экспериментатором изменчивость растений или шум, который может перекрыть ответную реакцию растения на изучаемый фактор в полевых опытах. Традиционно агрономы для снижения влияния внешней среды проводят опыт в нескольких повторностях, и чем их больше, тем точнее может быть опыт и больше вероятность обнаружить влияние изучаемого фактора, хотя этот прием экономически и практически весьма затратен.

В двадцатые годы прошлого столетия перед Министерством сельского хозяйства Англии возник вопрос как об удешевлении полевых испытаний сортов возделываемых культур на урожайность, так и о повышении точности получаемых данных. Это задание в 1923 году министерство поручило статисту Ротамстедской сельскохозяйственной станции Р.Фишеру, и он не только быстро разработал такой метод, но и успел к 1925 году опубликовать его в своей книге [6] в одной из глав под названием анализ вариансы. Эта методика оказалась весьма трудной для понимания [1-5] и не представлена ни как учебник, ни как руководство для исследователей по дисперсионному анализу. Это, скорее всего, отчет перед министерством и декларативная заявка на создание нового метода в статистике.

Впоследствии Р.Фишер писал, что для него это был бизнес. Может поэтому он преднамеренно при публикации метода усложнил процедуру вычисления конечных результатов. Хотя фактически, при его понимании, метод довольно прост, и весь анализ построен на разложении суммы квадратов отклонений экспериментальных дат от среднеарифметической величины всех дат. Так, например, для вычисления этой суммы квадратов им применяется сложнейшая процедура возведения в квадрат множества двух-, четырех- и более значных чисел. При наличии в то время только у немногих агрономов вычислительной машины в виде механического арифмометра «Феликс», даже вычисление путем

STUDIA UNIVERSITATIS MOLDAVIAE, 2015, nr.6(86)
Seria "Ştiinţe reale şi ale naturii" ISSN 1814-3237 ISSN online 1857-498X p.119-124

вращения ручки арифмометра только лишь суммы квадратов занимало несколько дней. Одно из таких усложнений вычисления видно на примере вычисления суммы квадратов отклонений S^2 по Р.Фишеру (1) и по методу обычной статистики на основе кривой нормального распределения по Пирсону (2).

$$S^2 = \sum x_i^2 - \frac{1}{N} (\sum x_i)^2 . \quad (1)$$

Вместо того, чтобы простейшим и очень легким вычитанием всех дат из среднеарифметической величины (2) получить сумму квадратов отклонений из одно- или двухзначных чисел:

$$S^2 = \sum (x_i - \bar{x})^2 . \quad (2)$$

Более того, формула (1) в скрытой форме представляет важнейшую взаимосвять между среднеквадратическим отклонением σ, средней арифметической величиной всех дат μ и среднеквадратической величиной всех дат φ:

$$\varphi^2 = \mu^2 + \sigma^2 \text{ или} \quad (3)$$

$$\sigma^2 = \varphi^2 - \mu^2 , \text{ где} \quad (4)$$

$$\sigma^2 = \frac{1}{N}\sum (x_i - \bar{x})^2 ; \ \varphi^2 = \frac{1}{N} \sum x_i^2 ; \ \mu^2 = (\frac{1}{N}\sum x_i)^2 . \quad (5)$$

Обратим внимание, что все три члена формулы (4) представлены в (5) делением на одно и то же число дат, равное N, и при замене хотя бы одного из них на так называемую степеннь свободы равенство нарушается. Подставив все три члена формулы (5) в формулу (4) и умножив их все на N, Р.Фишер получил формулу (1), с которой он начинает разложение дисперсии, сделав незаметно подмену дисперсии σ^2 на сумму квадратов отклонений всех дат от среднеарифметической величины этих дат s^2 путем умножения дисперсии на число дат N.

$$\sigma^2 * N = s^2 = \frac{1}{N} * N \sum (x_i - \bar{x})^2 , \ s^2 = \sum (x_i - \bar{x})^2 . \quad (6)$$

Компонент формулы (4) φ^2 после умножения на N стал называться суммой квадратов всех дат.

$$\varphi^2 * N = \frac{1}{N} * N \sum x_i^2 = \sum x_i^2 . \quad (7)$$

И наконец, квадрат среднеарифметической величины признака μ^2 после умножения на N поменял свое название на константу С.

Возможно такими манипуляциями Р.Фишер, не указывая этого в своих трудах, сконструировал формулу (1), с которой начинается разложение вариансы. При сопоставлении формулы (1) и формул (3,4) становится очевидным, что она описывает прямоугольный треугольник, построенный на основании, которым является катет, полученный извлечением квадратного корня из константы С. (Рис. 1). Во всех расчетах «разложения дисперсии» константа является неизменным основанием различных прямоугольных треугольников, у которых меняется гипотенуза, сумма квадратов дат и «число степеней свободы», на которые эта сумма квадратов дат делится, и полученный второй катет уже является искомым компонентом разложения дисперсии. Таким образом, все манипуляции ведутся с гипотенузой и степенями свободы.

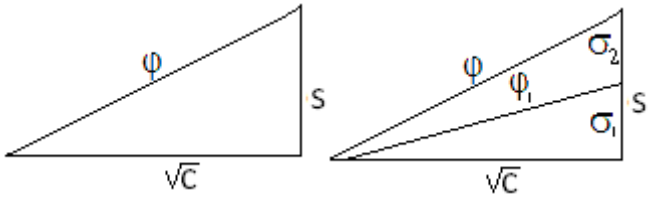

Рис.1. Схема разложения дисперсии по Р.Фишеру.

На рис.1 показана схема разложения дисперсии по Р.Фишеру, при которой константа С неизменна и при каждом вычислении она является вычитаемым членом последующего уравнения. Отсюда видно,

что изменяются только гипотенуза и степени свободы при расчете, и так, соответственно, находятся компоненты разложения общей дисперсии. Однако при этом всегда возникает ошибка в расчете, что и следует ожидать при такой аранжировке дат по Р.Фишеру. На рис. 2 представлен правильный способ разложения дисперсии по методу, описанному в данной статье.

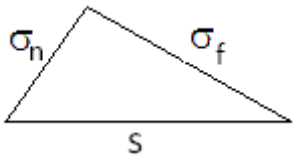

Рис.2. Принцип разложения дисперсии по методу, описанному в данной статье.

Вместе с тем, сама идея разложения дисперсии при полном понимании смысла происходящих вычислений и упрощении преднамеренно усложненной Р.Фишером процедуры вычисления весьма полезна в некоторых агрономических экспериментах. Это позволяет значительно уменьшить число повторностей делянок, чтобы отделить ответную реакцию растений на изучаемое воздействие от не контролируемых экспериментатором помех, вызываемых внешней средой. Этот метод называется в нашей стране дисперсионным анализом, поскольку анализ ведется не по количественным показателям изучаемых признаков растений, а по их изменчивости, или дисперсии.

В представленной Р.Фишером форме он оказался сложным для понимания и трудоемким для вычисления с неясно определенными конечными параметрами, такими как нулевая гипотеза и критерий Фишера [8]. Так, в самой книге Р.Фишера [6], где впервые упоминается этот метод, он теоретически не обоснован и его справедливость не доказана. В последующем Р.Фишер признал, что «его метод представляет удобную арифметическую аранжировку» экспериментальных дат [7]. Не дано техники расчета и применяемых при этом формул, а дана лишь конечная таблица получаемых при помощи математической перекомбинации [7] показателей, по которым оценивается опыт. Не указана важность центрированного представления дат повторностей при разложении вариансы. Хотя конечной целью всего дисперсионного анализа является доказательство нулевой гипотезы, по которой судят о том, действует ли изучаемый фактор на растения или нет, и она весьма туманно определена Р.Фишером. В последующем Р.Фишер был вынужден признать, что «нулевая гипотеза не доказывается, ни устанавливается, но скорее опровергается в ходе эксперимента. Каждый эксперимент, можно сказать, существует только для того, чтобы дать шанс для опровержения нулевой гипотезы» [8].

Доказательств метода и техники расчета нет и в книге А. Любищева [4], главного борца за его внедрение в практику сортоиспытания на сельскохозяйственных опытных станциях. Но главное, чего нет в этих книгах, так это то, что не изложен сам смысл дисперсионного анализа, его преимущества или недостатки, на основании которых можно узнать, в каких случаях его надо использовать, а в каких он бесполезен. Далее, традиционно, почти во всех руководствах по дисперсионному анализу дается лишь некий шаблон, по которому следует вести вычисления [5].

Учитывая этот пробел в литературе по дисперсионному анализу, в предложенной исследователям данной статье акцент сделан больше на смысловой части в облегченной математической форме, предназначенной дать агрономам и биологам более широкое понимание основы дисперсионного анализа, кратко изложенной ранее [2,3].

Дисперсионный анализа основан на математическом рассмотрении всех опытных данных как элементов единой кривой нормального распределения, центрированных вокруг единого центра рассеивания, за который принимается средняя арифметическая величина всей совокупности экспериментально полученных дат. Таким приемом, математически все количественные признаки становятся центрированными. За центр рассеивания или дисперсии принимается средняя арифметическая величина всех полученных экспериментальных дат, и все даты считаются членами одной совокупности, описываемой кривой нормального распределения.

Только общая дисперсия независимых друг от друга дат с их средней арифметической, равной нулю, или центрированных по отношению к их средней арифметической величине, равна сумме их частных дисперсий:

$\sigma^2 = \sigma_f^2 + \sigma_n^2$, где (8)

STUDIA UNIVERSITATIS MOLDAVIAE, 2015, nr.6(86)
Seria "Ştiinţe reale şi ale naturii" ISSN 1814-3237 ISSN online 1857-498X p.119-124

σ^2 – общая дисперсия всех дат от средней арифметической даты всей совокупности дат,

σ_f^2 – факториальная дисперсия, вызванная влиянием изучаемого фактора на растения,

σ_n^2 – шумовая дисперсия, вызванная шумом (noise) при воздействий внешней среды.

Формула (8) математически выражает закон Пифагора для прямоугольного треугольника, в котором общая дисперсия – это квадрат гипотенузы, а факториальная и шумовая дисперсии – это квадраты катетов этого прямоугольного треугольника. Отсюда разложить общую дисперсию на компоненты – это значит, что зная квадрат гипотенузы, найти квадраты катетов этого прямоугольного треугольника. Таким образом, даты должны иметь один общий центр рассеивания и должны быть центрированы. Тогда они будут принадлежать общей совокупности, описываемой кривой нормального распределения.

Чтобы дисперсионный анализ можно было провести вокруг единого центра рассеивания, необходимо перед постановкой полевого опыта так запланировать эксперимент, чтобы получился один общий центр рассеивания, и в итоге получить таблицу дат, в которой количество факторов и их повторностей было одинаково, или таблицу в виде квадрата. Наиболее оптимальная структура таблицы – это 5x5. Покажем пример составления центрированной таблицы дат для дисперсионного анализа.

Таблица 1

Исходные даты полевого опыта

Факторы y_j	Повторности x_i					$\sum y_j$	Средние факторов $\bar{y}_j$
	1	2	3	4	5		
1	67	67	55	42	47	278	55,6
2	98	96	91	66	77	428	85,6
3	60	69	50	35	43	257	51,4
4	79	64	81	70	63	357	71,4
5	90	70	79	88	72	399	79,8
$\sum x_i$	394	366	356	301	303		$\sum \bar{y}_j$
Средние шума $\bar{x}_i$	78,8	73,2	71,2	60,2	60,4		343,8
$\sum \bar{x}_i$	343,8						
Центр рассеивания всех дат	$\frac{1}{5}\sum \bar{y}_j = \frac{1}{5}\sum \bar{x}_i$ =343,8/5 = 68,76						

В конце расчета данных, представленных в таблице, убеждаемся, что сумма средних всех факторов и сумма средних всех дат шума равны одной величине 343,8. Это значит, что при делении на равное число колонок и строк мы получим единый центр рассеивания, которым является величина $\bar{x}$ = 68,76. Так, соответствующие им рассеивания или дисперсии действуют относительно единого центра рассеивания всех дат и они все центрированы. Теперь мы можем начать нахождение общей дисперсии и разложение ее на независимые друг от друга факториальную и шумовую части.

Таблица 2

Общие показатели дисперсионного анализа

N	Средние факториальные и повторностей x_i	Общая средняя средних $\bar{x}$	Отклонения от общей средней $x_i - \bar{x}$	Квадраты отклонений $(x_i - \bar{x})^2$	Общая дисперсия σ^2	Среднеквадратическое отклонение σ
1	55,6	68,76	-13,16	173,1856		
2	85,6	68,76	16,84	283,5856		
3	51,4	68,76	-17,36	301,3696		
4	71,4	68,76	2,64	6,9696		

5	79,8	68,76	11,04	121,8816	115,66	10,76
6	78,8	68,76	10,04	100,8016		
7	73,2	68,76	4,44	19,7136		
8	71,2	68,76	2,44	5,9536		
9	60,2	68,76	-8,56	73,2736		
10	60,4	68,76	-8,36	69,8896		
∑	687,6		-5E-14	1156,624		

Таблица 3

Факториальные показатели дисперсионного анализа

N	Средние факториальные $\bar{x}_f$	Общая средняя средних $\bar{x}$	Отклонения от общей средней $\bar{x}_f - \bar{x}$	Квадраты отклонений $(\bar{x}_f - \bar{x})^2$	Дисперсия факторов σ_f^2	Среднеквадратическое отклонение σ_f
1	55,6	68,76	-13,16	173,1856		
2	85,6	68,76	16,84	283,5856		
3	51,4	68,76	-17,36	301,3696	88,7	9,41
4	71,4	68,76	2,64	6,9696		
5	79,8	68,76	11,04	121,8816		
∑	343,8		0	886,992		

Таблица 4

Показатели шума дисперсионного анализа

N	Средние повторностей или шума $\bar{x}_n$	Общая средняя средних $\bar{x}$	Отклонения от общей средней $\bar{x}_n - \bar{x}$	Квадраты отклонений $(\bar{x}_n - \bar{x})^2$	Дисперсия шума σ_n^2	Среднеквадратическое отклонение σ_n
6	78,8	68,76	10,04	100,8016		
7	73,2	68,76	4,44	19,7136		
8	71,2	68,76	2,44	5,9536	26,96	5,19
9	60,2	68,76	-8,56	73,2736		
10	60,4	68,76	-8,36	69,8896		
∑	343,8		-5E-14	269,632		

Из таблиц 2,3 и 4 извлекаем искомые для нас дисперсии. Общая дисперсия равна $\sigma^2 = 115, 66$; факториальная дисперсия составляет часть общей дисперсии и равна $\sigma_f^2 = 88,7$ и дисперсия шума намного меньше и достигает величины $\sigma_n^2 = 26,96$. В сумме факториальная и шумовая дисперсии, при центрированном расположении дат, должны быть равны общей дисперсии. Никаких остатков дисперсии, обозначаемых в некоторых руководствах [5] по дисперсионному анализу как σ_z^2, не должно быть. Если они есть, то это значит, что данные не центрированы вокруг единого центра и схема опыта запланирована неправильно, либо дисперсионный анализ проведен с математическими погрешностями. Правильно проведенное разложение общей дисперсии позволяет обойтись без критерия Фишера и его нулевой гипотезы. Они в данном случае излишни и поэтому более наглядно степень влияния факториальной и шумовой дисперсий на результат опыта лучше выразить в процентах, приняв общую дисперсию за 100%, в соответствии с формулами, полученными преобразованием формулы (8):

$$\sigma^2 = \sigma_f^2 + \sigma_n^2, \quad \frac{\sigma^2}{\sigma^2} = \frac{\sigma_f^2}{\sigma^2} + \frac{\sigma_n^2}{\sigma^2}, \quad (1 = \frac{\sigma_f^2}{\sigma^2} + \frac{\sigma_n^2}{\sigma^2}) * 100\%. \qquad (9)$$

Подставим в формулу (9) полученные нами дисперсии из таблиц 2-4. Отношение факториальной дисперсии к общей равно:

$$\eta_f = \frac{\sigma_f^2}{\sigma^2} = \frac{88,6}{115,7} * 100 = 78,7\% , \quad (10)$$

отсюда доля факториальной дисперсии в общей составляет 78,7%. Это значительное ее превышение по сравнению с вкладом шумовой доли η_n, которая составляет

$$\eta_n = \frac{\sigma_n^2}{\sigma^2} = \frac{28,96}{115,7} * 100 = 23,3\% . \quad (11)$$

В случае если факториальная дисперсия равна дисперсии шума, то каждая из них составляет по 50% от общей дисперсии, тогда выделить из общей дисперсии факториальную дисперсию не представляется возможным. Но этого недостаточно, чтобы отвергнуть влияние изучаемого фактора. Это лишь предварительный вывод. Пока он гласит о недостаточной разрешающей способности опыта из-за высокого уровня шума внешней среды, не позволяющей выделить ответную или факториальную реакцию растения, либо о том, что необходимо уменьшить шум и полевой опыт заменить на лабораторный или провести его при контролируемых условиях в фитотроне.

ВЫВОД

С помощью теории нормального распределения проанализирован способ анализа вариансы, разработанный Р.Фишером. Разработан новый математически облегченный способ дисперсионного анализа для изучения степени влияния различных контролируемых факторов на растения в полевых агрономических и биологических исследованиях в неконтролируемых условиях внешней среды.

Литература:

1. БАБИЦКИЙ, А.Ф. Критерии для оценки биологических количественных признаков. В: *Современное состояние и пути развития популяционной биологии*. Материалы 10 Всероссийского популяционного семинара. (г. Ижевск, 17-22 ноября 2008 г.). Ижевск, 2008, с.9-12.
2. БАБИЦКИЙ, А.Ф. Как биологу понять смысл дисперсионного анализа. В: *Биологическая защита растений на пути инноваций*. Украинская научно-исследовательская станция карантина растений ИЗР НААН. Информационный бюллетень. Черновцы – Бояны, 2012, № 43, с.20-24.
3. БАБИЦКИЙ, А.Ф. Истоки создания дисперсионного анализа и его сущность. В: *Математическое моделирование в образовании, науке и производстве*. Материалы IX международной конференции «ММ 2015». (Тирасполь, 8-10 октября. 2015). Тирасполь, 2015, с.185-186.
4. ЛЮБИЩЕВ, А.А. *Дисперсионный анализ в биологии*. Москва: МГУ, 1986. 200 с.
5. ПЛОХИНСКИЙ, Н.А. *Биометрия*. Москва: МГУ, 1970. 358 с.
6. ФИШЕР, Р.А. Статистические методы для исследователей. Москва: Гостиздат, 1958. 267 с.
7. FISHER, R.A. The analysis variance is not mathematical theorem, but rather a convient method of arranging the arithmetic. Discussion to Statistics in agricultural research by Wichart. In: *Journal of the Royal Statistical Society*, Supplement, 1, 1934, p.26-61.
8. FISHER, R.A. *The design of experiments*. NY: Hafner Press, 1974, p.16.

Prezentat la 10.10.2015