17. ДИСПЕРСИОННЫЙ АНАЛИЗ: ОСНОВЫ

Развитие компьютерной техники и наличие стандартных пакетов программ по статистике (STATGRAPHICS, STATISTICA, SAS, LSMLMW и др.) делает возможным использование в научных исследованиях не только описательной статистики, корреляционного и регрессионного анализов, но и такого эффективного по разрешающей способности метода, как дисперсионный анализ.

17.1. Определение

В вариабельности признака отражается *интегрированный* вклад всех раздельно и совместно действующих факторов наследственности и среды. Эта общая *фенотипическая* изменчивость измеряется обычным способом - через стандартное отклонение и коэффициент изменчивости. Однако сила воздействия разных факторов неодинакова. Кроме того, влияние каждого фактора «заслоняется» действием всех остальных факторов. Вместе с тем, для эффективного разведения животных чрезвычайно важно знать в отдельности вклад каждого фактора в изменчивость признака.

Дисперсионный анализ - это (1) статистический метод обработки результатов наблюдений, зависящих от различных одновременно действующих факторов, (2) выбор наиболее важных факторов и (3) оценка их влияния (см. также [70, 72, 82, 93, 101]).

Исторически современный метод дисперсионного анализа (Analysis of Variance, ANOVA) разрабатывался, главным образом, применительно к задачам сельского хозяйства. Фундаментальная концепция метода была представлена в 1918-25 годах английским ученым Р. Фишером [119, 68].

Суть дисперсионного анализа заключается в проверке статистической значимости различия между средними (для групп или переменных). Эту проверку проводят с помощью разбиения общей суммы квадратов (SS_y) на компоненты*. Одна компонента обусловлена случайной ошибкой (т.е. внутригрупповой

* Более естественным называть рассматриваемый метод, как «анализ суммы квадратов» или «анализ варианс», но в силу традиции употребляют термин «дисперсионный анализ» (см.также [**122, 129÷131, 137, 138, 145**]).

изменчивостью) - её называют *суммой квадратов ошибки* или SS_e. Вторая компонента связана с различием средних значений (т.е. с межгрупповой изменчивостью). Её называют *суммой квадратов фактора* или SS_f. В общем:

$$SS_y = SS_f + SS_e.$$

Допустим, что необходимо изучить изменчивость признака, которая обусловлена несколькими факторами (например, фактор А – порода и фактор В – уровень кормления). В данном случае факториальную сумму квадратов (SS_f) может быть представлена суммой квадратов фактора А, фактора В и совместного влияния обоих факторов - АВ:

$$SS_f = SS_a + SS_b + SS_{ab}.$$

Тогда общая сумма квадратов есть:

$$SS_y = SS_a + SS_b + SS_{ab} + SS_e.$$

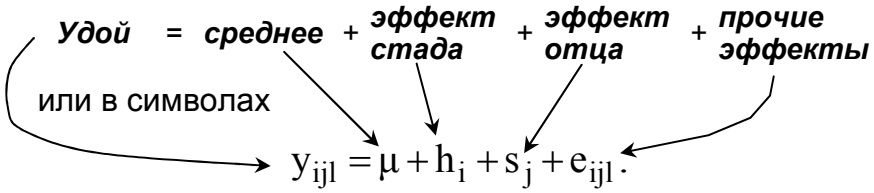
Дисперсионный анализ осуществляется при помощи процедуры *наименьших квадратов*, в основе которой метод *подбора констант* «method of fitting constants» (Gauss, 1809). Процедура наименьших квадратов минимизирует SS_e. Чем меньше SS_e, тем больше изучаемый признак подвержен влиянию контролируемых факторов, и тем лучше выявляется и познается изменчивость признака, возникающая под действием этих факторов.

С помощью дисперсионного анализа можно:

- выявить долю (степень) влияния различных факторов и их взаимодействие на изменчивость признака;
- определить статистическую значимость влияния изучаемых факторов;
- оценить компоненты фенотипической вариансы (ковариансы) рандомизированных факторов, в том числе генетических;
- получить наилучшие линейные несмещенные оценки градаций фиксированных факторов (или различий между градациями) и их статистическую значимость;
- установить величину криволинейной связи между факторами и анализируемым признаком.

17.2. Статистическая модель

Для проведения дисперсионного анализа важно точно определить статистическую модель. *Модель* - это уравнение, которое показывает, какие *независимые* переменные (факторы) влияют на зависимую *переменную* (признак). Например, удой дочери j-го быка в i-ом стаде можно записать так:



Если допустить, что средний удой по породе равен 4000 кг, эффект стада -200 кг, а племенная ценность отца +600 кг, то продуктивность первотелки можно записать так:

$$y_{ijl} = 4000 + (-200) + ½(+600) + 0 = 4100.$$

Племенную ценность отца умножают на ½, потому, что он передал дочери только половину своих генов.

Каждый включенный в модель фактор имеет не одно, а несколько (минимум два) значений, которые называют *уровнями* фактора, *классами* или *градациями*.

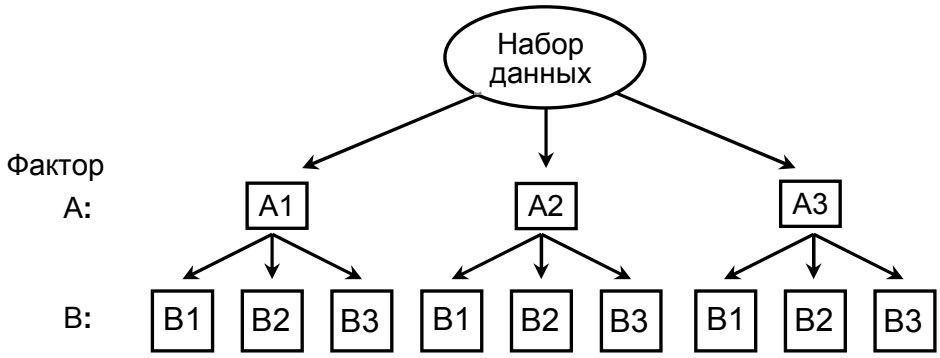
В биометрической модели факторы могут быть представлены как *фиксированные*, и как *случайные* (*рандомизированные*). Если уровни факторов точно установлены (например, три рациона, четыре сезона отела) или характеризуют определенное состояние (самцы, самки), или оценивают влияние отдельных градаций фактора из всех возможных (четыре конкретных стада из двадцати), то такие факторы являются фиксированными.

В биозоотехнических исследованиях могут быть факторы, уровни которых не являются точно фиксированными, или уровни которых являются следствием случайного выбора из большого числа градаций (например, 40 быков из 100), или которые вообще имеют все возможные случайные градации. Такие факторы называют случайными или рандомизированными. Под этим понимают только то, что случайными могут быть их разные уровни (случайные уровни некоторых факторов можно сделать фиксированными).

17.3. Типы моделей

В дисперсионном анализе возможны разные схемы или модели. По числу учитываемых факторов они могут быть *одно-, двух-, трехфакторные* и т.д.; в зависимости от структуры набора данных - *классификационные* или *иерархические* (*гнездовые*).

Классификационные модели. В моделях данного типа градации факторов качественные (дискретные) или «квантированные» (например, порода, пол, рацион, группировка типа «опыт-контроль»). Классификационными моделями анализируют данные со следующей структурой:



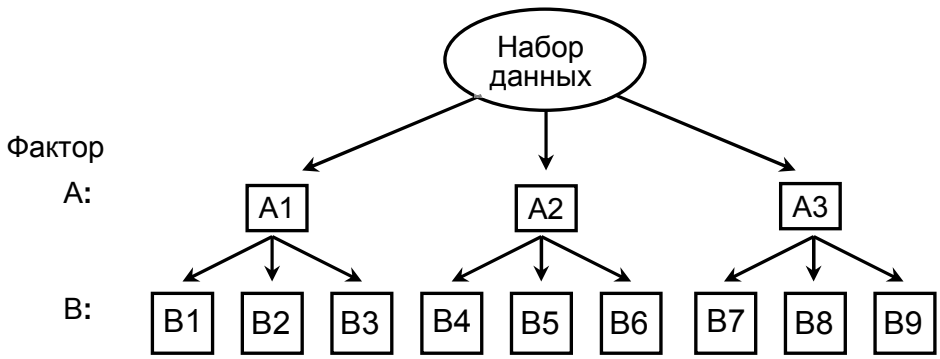
Классификационная модель:

$$y = \mu + A + B + e.$$

Модель может включать эффекты взаимодействия фактора A с фактором B, т.е. эффект A×B, например, взаимодействие «бык×стадо» («генотип×среда»):

$$y = \mu + A + B + A \times B + e.$$

Иерархические (гнездовые) модели. Моделями данного типа анализируют выборки со следующей структурой:



Иерархическая модель включает эффект фактора B внутри фактора A - B:A или B(A), например, быки внутри линии:

$$y = \mu + A + B:A + e$$

адекватно $$y = \mu + A + B(A) + e .$$

Выше отмечалось, что факторы могут быть фиксированными и случайными. В соответствии с этим различают *фиксированные, случайные* и *смешанные* модели.

Фиксированные модели используют для оценки эффектов выборочных градаций факторов. Выводы распространяют только на анализируемый набор данных.

Случайные модели используют для оценки компонентов общей вариансы. Выводы распространяют на всю популяцию.

Смешанные модели используют для одновременной оценки эффектов фиксированных факторов и варианс случайных факторов. На смешанных моделях основана процедура наилучшего линейного несмещенного прогноза генотипа животных (BLUP).

Если в число независимых переменных входят *непрерывные* количественные переменные, то такие статистические модели называют «*модели с регрессией*» или «*модели с ковариацией*», или «*модели с сопутствующими переменными*». Анализ по моделям, которые содержат как классификационные, так и непрерывные независимые переменные, называют *ковариантным*.

Схему (модель) дисперсионного анализа называют *полной*, если наблюдения имеются для каждой комбинации факторов (на пересечении строк и столбцов число наблюдений):

Фактор	A1	A2	A3
B1	10	10	10
B2	10	10	10
B3	10	10	10

Если число наблюдений для каждой комбинации факторов (в ячейках) равно между собой, то схему называют *сбалансированной (ортогональной)*. В таких случаях наилучшим является использование статистической процедуры ANOVA (Analysis of Variance - анализ вариансы).

При наличии пустых (без наблюдений) ячеек схему называют *неполной*:

Фактор	A1	A2	A3
B1	110	10	-
B2	-	105	100
B3	210	10	10

Если число наблюдений для комбинаций факторов различно, то такие схемы называют *несбалансированными (неортогональными)*.

Ситуации, которые описываются неполными несбалансированными моделями, наиболее типичны в животноводстве. В таких случаях использовать процедуру ANOVA нельзя. Для неполных несбалансированных моделей необходимо использовать статистическую процедуру GLM (General Linear Models - обобщенные линейные модели).

Примеры биометрических моделей дисперсионного анализа:

Обозначим:

y, y_1, y_2 - зависимые переменные;

A, B, C - фиксированные независимые классификационные переменные (главные факторы);

X_1, X_2 - фиксированные независимые непрерывные переменные (коварианты);

s - рандомизированная независимая переменная.

Возможные модели:

$y = A+e$ - однофакторная модель;

$y = A+B+C+e$ - модель главных факторов;

$y = A+B+A\times B+e$ - факторная модель (с взаимодействием);

$y = A+B(A)+C(BA)+e$ - гнездовая модель;

$y_1\ y_2 = A+B+e$ - мультивариантная модель;

$y = A+X_1+e$ - модель с регрессией;

$y = A+X_1(A)+e$ - модель с частной регрессией;

$y = A+X_1+X_1\times A+e$ - модель с взаимодействием главного фактора и регрессионной переменной;

и множество других моделей.

Пример расширенной модели:

$$y = \mu + A + B + A \times B + C(A) + X_1 + X_1 \times X_1 + X_2(A) + s + e.$$

Данная модель включает:

- общее среднее μ;
- эффекты фиксированных факторов A, B, C;
- взаимодействие между A×B;
- эффект фактора C внутри фактора A (гнездовой эффект);
- вероятные эффекты количественных переменных, причем может быть линейная X_2 и квадратичная $X_1+X_1 \times X_1$ формы зависимости;
- гнездовой эффект непрерывной переменной внутри фиксированного фактора $X_2(A)$;
- эффект рандомизированного фактора **s** (например, отца);
- остаточные неучтенные рандомизированные эффекты e, показывающие изменчивость зависимой переменной Y, которую модель не может объяснить (усредненная характеристика индивидуальной внутригрупповой изменчивости или остаточная вариация, или ошибка).

17.4. Нулевая гипотеза

В дисперсионном анализе гипотезу H_0, как правило, формулируют в терминах равенства групповых средних. Например, если исследуют удой коров четырех пород, то формулировка гипотезы H_0 будет следующей: «Средние удои коров четырех пород достоверно не различаются».

Фактически, гипотеза H_0 состоит в допущении, что *изменчивость средних* по породам не значима (не существенна, статистически не достоверна). Если на уровне α гипотеза H_0 будет отклонена, то говорят, что изменчивость средних значима и между породами имеют место статистически значимые различия.

При проверке значимости используют предложенной Фишером F-критерий. Если рассчитанное значение F-критерия будет меньше табличного критического значения (находят по табл. А.10, А.11 Приложения А), то гипотезу H_0 принимают. Влияние рассматриваемого фактора (порода) нет основания считать значимым. В противном случае гипотезу H_0 отвергают и утверждают, что существуют различия между эффектами отдельных градаций фактора (между породами). Чтобы выяснить,

какие именно градации (породы) отличаются друг от друга, оценивают значимость различий между градациями (попарное сравнение средних по породам).

Таким образом, дисперсионный анализ - это метод исследования, основанный на сравнительном изучении рассеяния случайных величин. Поэтому, чем больше число наблюдений, тем стабильнее оценки параметров, тем бóльшая вероятность обнаружения тех или иных закономерностей, свойственных изучаемому явлению. Влияние неконтролируемых причин при этом выравнивается и теряет силу. Это происходит в результате действия закона больших чисел: совместное влияние большого числа случайных факторов приводит к результату, почти не зависящему от случая.

17.5. Стандартный анализ

Рассмотрим простую однофакторную модель:

$$y_{ij} = \mu + a_i + e_{ij},$$

где y_{ij} - вес j-го животного из i-ой группы; μ - общее среднее по выборке из n животных; a_i - эффект, обусловленный влиянием i-го уровня фактора (например, рациона, при $i=1,...,p$,); e_{ij} - случайная ошибка, специфическая для j-го животного ($j=1,...,n_i$, причем n_i - число животных в i-ой группе).

В одних случаях a_i может быть фиксированным эффектом, в других - рандомизированным со средней равной 0 и вариансой σ_a^2.

34. Символьное представление исходных данных

	Градации фактора А			Общая сумма
	$a_1 \ldots$	$a_i \ldots$	a_p	
Живая масса животных (повторности)	y_{11} y_{12} $\vdots$	y_{i1} y_{i2} $\vdots$	y_{p1} y_{p2} $\vdots$	
Сумма по группам	$\sum_j y_{1j} = y_{1.}$	$\sum_j y_{ij} = y_{i.}$	$\sum_j y_{pj} = y_{p.}$	$\sum_i \sum_j y_{ij} = y_{..}$

В табл. 34 представлены исходные данные в символьном виде, а в табл. 35 дан стандартный алгоритм дисперсионного анализа по однофакторной модели.

35. Алгоритм дисперсионного анализа по однофакторной модели

Источник изменчивости	df	SS	MS	E(MS)
Общая	n	$SS_y = \sum_i \sum_j y_{ij}^2$	-	-
Среднее	1	$SS_\mu = y_{..}^2 / n$	-	-
Общая скорректированная	n-1	$SS_y^* = SS_y - SS_\mu$	-	-
Между градациями фактора А	p-1	$SS_a = \sum_i \frac{y_{i.}^2}{n_i} - SS_\mu$	$MS_a = \frac{SS_a}{df_a}$	$\sigma_e^2 + k\sigma_a^2$
Внутри градаций фактора А (ошибка)	n-p	$SS_e = SS_y - \sum_i \frac{y_{i.}^2}{n_i}$	$MS_e = \frac{SS_e}{df_e}$	σ_e^2

Примечание. **df** – число степеней свободы; **SS** - сумма квадратов;
MS – средний квадрат; **E(MS)** – математическое ожидание среднего квадрата.

Если фактор А фиксированный, то нулевая гипотеза имеет вид: $H_0 : a_1 = ... = a_i = ... = a_p = 0$. Оценивают μ и эффекты a_i из отношений ($\bar{y}_i$ - среднее по i-ой группе):

$$\hat{\mu} = \frac{1}{p}\left(\frac{y_{1.}}{n_{1.}} + ... + \frac{y_{i.}}{n_{i.}} + ... + \frac{y_{p.}}{n_{p.}}\right) = \frac{1}{p}\sum \bar{y}_i ,$$

$$\hat{a}_i = \bar{y}_i - \hat{\mu} .$$

Ошибки оценок $\hat{\mu}$, $\hat{a}_i$ и $\hat{\mu}_i \,(= \hat{\mu} + \hat{a}_i)$:

$$m_{\hat{\mu}} = \sqrt{\frac{\hat{\sigma}_e^2}{n}}; \quad m_{\hat{a}_i} = \sqrt{\frac{\hat{\sigma}_e^2}{n_i} + \frac{\hat{\sigma}_e^2}{n}} = \hat{\sigma}_e \sqrt{\frac{1}{n_i} + \frac{1}{n}} \quad \text{и} \quad m_{\hat{\mu}_i} = \sqrt{\frac{\hat{\sigma}_e^2}{n_i}} .$$

Если фактор А рандомизированный, то нулевая гипотеза имеет вид: $H_0 : \sigma_a^2 = 0$. Оценивают остаточную (σ_e^2) и факториальную (σ_a^2) варианси посредством приравнивания средних квадратов к их математическим ожиданиям:

$$MS_a = \sigma_e^2 + k\sigma_a^2 \quad \text{и} \quad MS_e = \sigma_e^2 .$$

Тогда

$$\hat{\sigma}_e^2 = MS_e \quad \textbf{и} \quad \hat{\sigma}_a^2 = (MS_a - \hat{\sigma}_e^2)/k .$$

k - средневзвешенное число животных в градациях фактора А:

$$k = \frac{n - \sum_i n_i^2}{p - 1}.$$

В конечном итоге нужно узнать: (1) отличается ли $\hat{\sigma}_a^2$ от нуля, и (2) если отличается, то оценить долю (силу) влияния изучаемого фактора в общей вариации, т.е. оценить величину

$$\hat{r}_w = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2},$$

которую называют *коэффициентом внутриклассовой корреляции*.

Проверка нулевой гипотезы. Если допустить, что ошибки наблюдений распределены нормально со средней равной нулю и вариансой $\hat{\sigma}_e^2$, то проверку нулевой гипотезы сводят к рассмотрению соотношения:

$$F = \frac{MS_a}{MS_e} \quad \text{или} \quad F = \frac{MS_e}{MS_a}$$

(в общем, бóльший средний квадрат делят на меньший).

Процедура тестирования H_0 одинакова для фиксированных и рандомизированных моделей. Она включает:

- выбор уровня значимости α;
- выбор табличного критического значения - $F_{\alpha;\, df_1, df_2}$;
- расчет фактического значения F-критерия;
- принятие решения относительно H_0.

При принятии решения $F_{факт}$ сравнивают с $F_{табл}$.

Если $F_{факт} < F_{табл.}$ при уровне значимости α и числе степеней свободы df_1 для бóльшего и df_2 для меньшего средних квадратов, то гипотезу H_0 принимают. Различия между средними признают незначимыми на уровне α, влияние изучаемого фактора считают недоказанным, т.е. вся имеющаяся изменчивость признака обусловлена случайными причинами.

Если $F_{факт} \geq F_{табл.}$, то гипотезу H_0 отклоняют, т.е. по крайней мере одно из *p* сравниваемых средних статистически значимо отличается от остальных.

Проверка гипотезы H_0 подобна принципу «презумпции невиновности» в юриспруденции: обвиняемого считают невиновным до тех пор, пока его вина не будет доказана. При наличии улик, подтверждающих, что подозреваемый совершил преступление, его «невиновность» отвергают.

Уровень значимости α есть вероятность отвергнуть проверяемую гипотезу в том случае, когда она верна, т.е. совершить ошибку первого рода. Если же эта гипотеза неверна, а верна некоторая гипотеза H_1, то имеется возможность совершить с определенной вероятностью β ошибку второго рода: принять неверную нулевую гипотезу (H_0).

17.6. Результаты анализа

Пусть имеются данные по 18 поросятам, для которых определена следующая статистическая модель:

$$y_{ijk} = \mu + s_i + r_j + e_{ijk},$$

где y_{ijk} - живая масса k-го поросенка (полусибса), получавшего j-ый рацион и имевшего i-го отца; μ - общее среднее при равной численности градаций; s_i - эффект i-го отца (фиксированный); r_j - эффект j-го рациона (фиксированный); e_{ijk} - случайная ошибка.

После обработки данных по этой модели практически все компьютерные программы выдают следующую таблицу:

36. Результаты дисперсионного анализа (программа SAS, LSMLMW)

SOURSE [1]	df [2]	SS [3]	MS [4]	F [6]	Pr>F [7]
TOTAL	17	77,1	-	-	-
MODEL	3	20,8	6,9	1,73	0,207
S	2	15,7	7,8	1,95	0,179
R	1	9,7	9,7	2,41	0,142
ERROR	14	56,3	4,0 [5]	-	-
R-SQUARE = 0,27 [8]					

Комментарии к табл. 36:

1. Sourse - источник изменчивости: TOTAL - общая, MODEL - *все* учтенные в модели факторы, S - первый и R - второй главные факторы, ERROR - неучтенные факторы (ошибка).
2. df - число степеней свободы - это число наблюдений или градаций фактора *функционально* не связанных друг с другом.
3. SS - сумма квадратов для зависимой переменной, которая раскладывается на общую (SS_y), факторные (SS_s, SS_r), модели (SS_m, $SS_m = SS_s + SS_r$) и остаточную (ошибку, SS_e).
4. MS - средний квадрат - это сумма квадратов, деленная на число степеней свободы.
5. MSe - средний квадрат ошибки - это оценка остаточной вариансы (σ_e^2), варианса ошибки. *Корень квадратный из остаточной вариансы есть стандартное отклонение зависимой переменной*.
6. F - критерий Фишера. Это частное от деления среднего квадрата для фактора, включенного в модель, на средний квадрат ошибки (или наоборот). В общем случае больший средний квадрат делят на меньший. Критерий Фишера является тестом нулевой гипотезы. Нулевая гипотеза содержит утверждение, что включенные в модель факторы не оказывают достоверного влияния на зависимую переменную. Если нулевая гипотеза остается в силе, то вся изменчивость признака обусловлена случайными (не включенными в модель) факторами.
7. Pr>F (или p-level, или p-valio, или p-value) - уровень значимости или ошибки; тот минимальный уровень, на котором можно отвергнуть гипотезу; вероятность неправильного отвержения гипотезы, когда она верна. Значение Pr>F=0,142 для R – это большое число. На основании его нельзя отвергнуть гипотезу о равенстве средних по живой массе поросят, получавших разные рационы. Собранные данные не дают оснований говорить о том, что рацион влиял на живую массу поросят.
8. R-Square - коэффициент детерминации (R^2); измеряет, какая доля общей изменчивости зависимой переменной объясняется факторами, включенными в модель (после корректировки на среднюю). Это отношение суммы квадратов по модели к общей скорректированной на μ сумме квадратов ($R^2 = SS_m / SS_y^*$ или $R^2 = 1 - SS_e / SS_y^*$). Варьирует от 0 до 1 (от 0 до 100%). Чем больше значение R^2, тем лучше модель описывает зависимую переменную.

Различают частные коэффициенты детерминации, которые показывают силу влияния *каждого изучаемого фактора* на изменчивость признака. Например, $R_s^2 = SS_s / SS_y^*$, в отечественной литературе - η^2 (*эта*-квадрат).

Корень квадратный из η^2 (или R^2) называют корреляционным отношением; может измерять степень криволинейности связи между признаком и фактором(-ами).

Значимость различий по F-критерию не означает, что каждый эффект градации фактора отличается от другого. Между некоторыми градациями может *не быть* достоверных различий. Результаты выводятся в таблицах LS-оценок констант (табл. 37).

37. Оценки параметров по 2-х факторной модели (константы наименьших квадратов, LS)

PARAMETER	ESTIMATE [9]	T for H0 [10]	Pr>|H| [11]	STD [12]
По программе SAS ($\hat{s}_3 = 0, \quad \hat{r}_2 = 0$)				
INTERCEPT (μ)	5,3	6,3	0,0001	0,84
S 1	-0,5	-0,3	0,7361	1,34
2	1,7	1,5	0,1632	1,18
3	0,0	…	…	…
R 1	-1,6	1,6	0,1425	1,04
2	0,0	…	…	…
По программе LSML ($\sum \hat{s}_i = 0, \quad \sum \hat{r}_j = 0$)				
μ	4,9	-	-	0,49
S 1	-0,9	-	-	0,76
2	1,3	-	-	0,67
3	-0,4	-	-	0,73
R 1	-0,8	-	-	0,52
2	0,8	-	-	0,52

Комментарии к табл. 37:

9. ESTIMATE - оценки LS-методом средних значений и эффектов, включенных в модель факторов (константы LS).
10. T for H0 - критерий Стьюдента для проверки нулевой гипотезы (параметр = 0).
11. Pr>|H| (или p-level, или p-valio, или p-value) - уровень значимости или ошибки; вероятность неправильного отвержения гипотезы, когда она верна.
12. STD - стандартная ошибка оценки относительно истинного значения параметра.

17.7. Допущения и оценки

В дисперсионном анализе очень важно четко определить не только биометрическую модель, но и знать допущения модели, принятые в той или иной компьютерной программе.

Пусть имеется однофакторная модель:

$$y_{ij} = \mu + a_i + e_{ij}$$
$$= \mu_i + e_{ij},$$

где y_{ij} - j-ое значение признака в i-ой градации фактора A; μ - общее среднее при равном числе наблюдений в каждой градации; μ_i - среднее по i-ой градации фактора A; a_i - фиксированный эффект i-ой градации фактора A, выраженный как отклонение от общего среднего μ; e_{ij} - случайные эффекты неучтенных факторов, которые предполагают независимыми (случайная ошибка).

Статистические модели обычно содержат больше неизвестных параметров, чем имеется однозначных решений. LS-методом можно получить однозначные оценки только для μ_i

$$\hat{\mu}_i = \bar{y}_{i.} = \sum_{i=1}^{n_i} y_{ij} / n_i,$$

где $i=1, 2,..., n_i$ - число наблюдений в i-ой градации фактора.

Однако для μ и a_i однозначных решений нет. Чтобы получить оценки этих параметров необходимо принять дополнительные условия, т.н. *допущения* (параметры или комбинации параметров являются оцениваемыми, если оценки однозначны независимо от каких-либо допущений).

В разных компьютерных программах используют различные допущения. В частности:

- оценка эффекта последней градации фактора равна 0, $\hat{a}_p = 0$, (допущение, заложенное в программе SAS);
- сумма оценок эффектов по всем градациям фактора равна нулю, $\sum \hat{a}_i = 0$, (допущение, заложенное в программе LSMLMW);
- взвешенная сумма оценок эффектов по всем градациям фактора равна нулю, $\sum w_i \hat{a}_i = 0$, (может быть использовано в программе LSMLMW).

Ниже дан пример того, как различные допущения влияют на оценки параметров, но не на их оценочные функции.

Пусть имеется пять наблюдений, распределенных по двум градациям фактора А (n=5):

Фактор:	a_1	a_2	Среднее
Значения:	1 2 3	4 6 -	
Сумма: n_i : $\bar{y}_{1.}$:	6 3 2	10 2 5	$\bar{y}_{..}$ не взвешенное = (2+5)/2=3,5 $\bar{y}_{..}$ взвешенное = (6+10)/5 = 3,2

Различные допущения относительно оцениваемых параметров приводят к различным решениям и, следовательно, к различным оценкам параметров:

Допущения	Оценки	Оценочные функции
1. $\hat{a}_2 = 0$	$\hat{\mu}$ = 5 $\hat{a}_1$ = -3 $\hat{a}_2$ = 0	- $\hat{\mu} + \hat{a}_1 = \mu_1 = 2$ $\hat{a}_1 - \hat{a}_2 = -3$
2. $\sum \hat{a}_i = 0$	$\hat{\mu}$ = 3,5 $\hat{a}_1$ = -1,5 $\hat{a}_2$ =+1,5	- $\hat{\mu} + \hat{a}_1 = \mu_1 = 2$ $\hat{a}_1 - \hat{a}_2 = -3$
3. $\sum n_i \hat{a}_i = 0$	$\hat{\mu}$ = 3,2 $\hat{a}_1$ = -1,2 $\hat{a}_2$ =+1,8	- $\hat{\mu} + \hat{a}_1 = \mu_1 = 2$ $\hat{a}_1 - \hat{a}_2 = -3$

Ранги эффектов (различия между эффектами) не зависят от принятых допущений. Важно знать, что суммы квадратов и оценки компонент фенотипической вариансы также не зависят от допущений биометрической модели.

17.8. R( )-запись

Чтобы показать, что содержит сумма квадратов (отклонений) конкретного типа дисперсионного анализа используют способ, называемый R( )-записью. R( )-запись указывает на ту часть суммы квадратов, которая объясняется моделью. Знак R обозначает

редукцию (reduction) суммы квадратов, обусловленных моделью, а в скобках указана модель. Так, для модели

$$y_{ij} = \mu + a_i + e_{ij}$$

сумму квадратов записывают, как $R(\mu, a)$, а модели

$$y_j = \mu + e_j$$

соответствует $R(\mu)$.

Если необходимо показать, что эффект μ элиминирован, то это записывают так: $R(a|\mu)$. Знак (|) читается как «дано». Корректировку на среднее суммы квадратов по фактору А записывают так:

$$R(a|\mu) = R(\mu, a) - R(\mu).$$

R( )-запись для 2-х факторных моделей:

- классификационная

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk} \rightarrow R(\mu, a, b, ab);$$

- иерархическая

$$y_{ijk} = \mu + a_i + b_{ij} + e_{ijk} \rightarrow R(\mu, a, b:a).$$

Чтобы получить интересующие суммы квадратов, модели редуцируют до

$$y_{ij} = \mu + a_i + e_{ij} \quad \text{и}$$

$$y_i = \mu + e_i.$$

Эффект взаимодействия после элиминации μ, a_i, b_j записывают так:

$$R(ab|\mu, a, b) = R(\mu, a, b, ab) - R(\mu, a, b),$$

где $R(\mu, a, b)$ - сумма квадратов для модели $y_{ijk} = \mu + a_i + b_j + e_{ijk}$.

$$R(a|\mu, b) = R(\mu, a, b) - R(\mu, b),$$

где $R(\mu, b)$ - сумма квадратов для модели $y_{ijk} = \mu + b_i + e_{ijk}$.

Такое разложение суммы квадратов является однозначным.

Редукция зависит от ограничений:

$$R'(a|\mu, b, ab) = R(\mu, a, b, ab) - R'(\mu, b, ab).$$

Апостроф «'» означает, что данная сумма квадратов находится в зависимости от используемых ограничений. Поэтому

разложение суммы квадратов не является однозначным. Причина заключается в том, что сумма квадратов для модели

$$y_{ijk} = \mu + b_j + ab_{ij} + e_{ijk}$$

не однозначна - взаимодействие учитывается, но не учитываются оба главных фактора. Поэтому применять надо такие решения, которые соответствуют ограничениям.

17.9. Типы сумм квадратов

Различают 4 типа разложения суммы квадратов: SS1, SS2, SS3 и SS4*.

Тип SS1. Называют также *последовательной* суммой квадратов. Зависит от порядка эффектов в модели. Каждый последующий эффект скорректирован только на предыдущие эффекты. Не учитывает эффекты, которые записаны за ним. В «чистом виде» оценивается только последний эффект модели. Является аддитивным по отношению к общей сумме квадратов. Гипотезы SS1 не зависят от предшествующих факторов, но зависят от последующих.

Пример: модель $y = A + B + A \times B$

Фактор	Тип SS1
A	$R(a \mid \mu)$
B	$R(b \mid \mu, a)$
A×B	$R(ab \mid \mu, a, b)$

Используют для:

- сбалансированных моделей при заданной последовательности без учета взаимодействия;
- иерархических моделей с определенной последовательностью эффектов;
- полиномиальных регрессионных моделей с определенной последовательностью.

Тип SS2. Дает частное разложение суммы квадратов. Не зависит от порядка эффектов в модели. Каждый эффект скорректирован на все остальные эффекты (кроме взаимодействия и

* В компьютерной программе SAS для того, чтобы быть уверенным - каким гипотезам отвечает статистическая модель и типу суммы квадратов, необходимо задать опцию E и соответствующий номер типа суммы квадратов.

гнездовых эффектов). Данный тип суммы квадратов *не обязательно* аддитивный по отношению к общей сумме квадратов модели.

Используют для:

- сбалансированных моделей;
- моделей с главными эффектами;
- регрессионных моделей.

Примеры:

Модель: $y = A + B + C$		Модель: $y = A + B + A \times B$	
Фактор	Тип SS2	Фактор	Тип SS2
A	$R(a \mid \mu, b, c,)$	A	$R(a \mid \mu, b)$
B	$R(b \mid \mu, a, c)$	B	$R(b \mid \mu, a)$
C	$R(c \mid \mu, a, b)$	A×B	$R(a \times b \mid \mu, a, b)$
Модель: $y = A + B(A) + C(AB)$		Модель: $y = X + X^2$	
Фактор	Тип SS2	Фактор	Тип SS2
A	$R(a \mid \mu)$	X	$R(x \mid \mu, x^2)$
B(A)	$R(b(a) \mid \mu, a)$	X^2	$R(x^2 \mid \mu, x)$
C(AB)	$R(c(ab) \mid \mu, a, b(a))$	-	-

Типы SS3 и SS4. Дают частное разложение суммы квадратов. Не зависят от порядка эффектов в модели. Каждый эффект скорректирован на все остальные эффекты. Если модель полная (нет пустых клеток) , то SS3=SS4. Если в модели нет эффекта взаимодействия, то SS2=SS3=SS4. Если есть пустые клетки и в модель включен гнездовой эффект, то SS3≠SS4. Типы SS3 и SS4 *обычно не аддитивны* к общей сумме квадратов. SS4 используют для неполных, несбалансированных моделей.

Пример: модель $y = A + B + A \times B$

Фактор	Тип SS3 и SS4
A	$R(a \mid \mu, b, ab)$
B	$R(b \mid \mu, a, ab)$
A×B	$R(ab \mid \mu, a, b)$

Обобщение. Модель:

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}$$
$$= \mu_{ij} + e_{ijk}.$$

Общая сумма квадратов для модели - $R(a, b, ab \mid \mu)$.

38. R( )-запись для разных типов суммы квадратов

Фактор	d.f.	Тип суммы квадратов		
		SS1	SS2	SS3, SS4
A	a-1	$R(a\|\mu)$	$R(a\|\mu,b)$	$R'(a\|\mu,b,ab)$
B	b-1	$R(b\|\mu,a)$	$R(b\|\mu,a)$	$R'(b\|\mu,a,ab)$
A×B	(a-1)(b-1)	$R(ab\|\mu,a,b)$	Как и SS1	Как и SS1

Отметим, что сумма квадратов R'( ) зависит от ограничений, принятых в анализе. Пусть:

$$\sum_i a_i = \sum_j b_j = \sum_i ab_{ij} = \sum_j ab_{ij} = 0,$$

тогда $R'(a|\mu,b,ab) = R(\mu,a,b,ab) - R'(\mu,b,ab).$

17.10. Сила влияния и корреляционное отношение

Одной из задач дисперсионного анализа является определение силы влияния фактора на изменчивость анализируемого признака. Показатель *силы влияния* рассчитывают из отношения суммы квадратов по j-му фактору к общей сумме квадратов:

$$\eta_j^2 = \frac{SS_j}{SS_y^*}.$$

η^2 (*эта*-квадрат) – это (1) долю общей изменчивости, которая приходится на факториальную; (2) вклад изучаемого фактора в общую изменчивость признака.

В основу показателя силы влияния положен принцип аддитивности:

$$\eta_1^2 + \eta_2^2 + ... + \eta_j^2 + ... + \eta_e^2 = 1.$$

Равенство выполнимо только для последовательной суммы квадратов (тип SS1). Для остальных типов суммы квадратов показатель силы влияния может быть рассчитан лишь приближенно:

$$\eta_j^2 = \frac{SS_j}{SS_1 + SS_2 + ... + SS_j + ... + SS_e}.$$

В биозоотехнических исследованиях часто имеет место *криволинейный* тип связи между переменными. Например,

изменения живой массы, продуктивности, плодовитости и т.д. характеризуются четкой криволинейной сопряженностью с возрастной динамикой животных. Для измерения тесноты связи такого типа используют *корреляционное отношение*, η, которое есть корень квадратный из показателя силы влияния:

$$\eta = \sqrt{\eta^2}\,.$$

Корреляционное отношение выражают десятичной дробью (от 0 до 1) или в процентах (от 0 до 100%). Чем больше η, тем сильнее связь между переменными. Корреляционное отношение не может быть меньше нуля, поэтому оно не определяет направление связи («+» или «-»).

Как и коэффициент регрессии, η имеет два значения: η_{xy} и η_{yx}, причем $\eta_{xy} \neq \eta_{yx}$. Это свойство позволяет выявить *неравнозначность* связи между коррелирующими переменными.

Коэффициент парной корреляции r всегда меньше η. Корреляционное отношение может указывать на наличие связи даже в тех случаях, когда коэффициент корреляции близок к нулю (что бывает при сильной криволинейности связи).

В принципе, все обычно наблюдаемые связи в той или иной степени криволинейны. Сопоставление η^2 и r^2 позволяет определить меру их криволинейности (L):

$$L = \eta^2 - r^2\,.$$

Чем больше значение L, тем менее пригоден показатель r для суждения о силе связи. При $\eta^2 = r^2$ связь принимает линейный характер.

Порог криволинейности, при превышении которого связь можно не считать прямолинейной:

$$F_L = \frac{(\eta^2 - r^2)(n-p)}{(1-\eta^2)(p-2)} \geq F_{\alpha;\, df_1, df_2}\,,$$

где $df_1 = p - 2$ и $df_2 = n - p$ (p – число градаций фактора).

17.11. Внутриклассовая корреляция

По биометрическим моделям с рандомизированными факторами можно рассчитать коэффициент *внутриклассовой корреляции* - статистика, которую часто используют в исследованиях по количественной генетике.

Сходство между родственными особями (полусибсы, сибсы), можно рассматривать и как *подобие* особей одной и той же группы, и как *различие* между особями разных групп. Чем больше подобие *внутри* групп, тем сильнее различие *между* группами.

Степень сходства можно определить отношением межгрупповой вариансы к общей вариансе. Межгрупповая варианса отражает ту долю изменчивости, которая является общей для членов одной и той же группы, а это не что иное, как *ковариация* членов групп.

Если пары переменных X и Y действительно являются парными выборками с общей переменной, то любую из них можно представить моделью

$$X_{ij} = \mu + a_i + e_{ij},$$

где X_{ij} - наблюдаемое значение для j-го члена i-ой пары или группы; μ - общее среднее; a_i - эффект i-ой группы (рандомизированный со средним 0 и вариансой σ_a^2); e_{ij} - случайная ошибка (рандомизированная со средним 0 и вариансой σ_e^2).

Корреляция между первым и вторым членами пары есть

$$r_w = \frac{\mathrm{Cov}(X_{i1}, X_{i2})}{\sqrt{\mathrm{Var}(X_{i1})\,\mathrm{Var}(X_{i2})}}.$$

Ковариансa между отдельными наблюдаемыми значениями членов группы (совокупность этих значений называют также классом) отлична от нуля и равна вариансе σ_a^2 величин a_i:

$$\mathrm{Cov}(X_{i1}, X_{i2}) = \mathrm{Cov}[(\mu + a_i + e_{i1}),(\mu + a_i + e_{i2})] =$$
$$= \mathrm{Cov}(a_i, a_i) = \sigma_a^2.$$

Эту ковариансу называют *внутриклассовой*.

Относительно варианс имеем:

$$\mathrm{Var}(X_{i1}) = \mathrm{Var}(X_{i2}) = \\ = \mathrm{Var}(\mu + a_i + e_{ij}) = \sigma_a^2 + \sigma_e^2$$

Подставляя окончательные значения ковариансы и варианс в формулу для r_w получим

$$r_w = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}.$$

Это есть *коэффициент внутриклассовый корреляции* (символ «w» - от слова *within* - внутри). Коэффициент служит мерой связи между двумя любыми значениями одного класса; его величина одинакова для всех классов (всех градаций фактора).

Внутриклассовая корреляция имеет место там, где переменные повторяются не в парах, а в больших группах. Поэтому она полезна как мера связи между особями внутри групп, например, полусибсов.

Группа полусибсов представляет потомство от спаривания одной особи со случайно выбранной группой животных другого пола (от каждого спаривания один потомок). Среднее генетическое такой группы (a_i) равно ½ племенной ценности (BV_i) их общего родителя (a_i = ½BV_i). Коварианса между членами группы соответствует варианс истинных средних групп и равна, следовательно, варианс ½BV общих родителей. Это составляет ¼ часть аддитивной генетической вариансы (σ_A^2):

$$\mathrm{Cov}(a_i, a_i) = \sigma_a^2 = (1/4)\sigma_A^2.$$

Степень сходства между полусибсами выражают через внутриклассовую корреляцию. Эта корреляция отражает межклассовую вариансу, то есть ковариацию, которая является частью общей вариансы. Поэтому для полусибсов

$$r_w = \frac{1}{4}\frac{\sigma_A^2}{\sigma_y^2},$$

где σ_y^2 - общая фенотипическая варианса ($=\sigma_P^2$).

Величина r_w тем больше, чем сильнее взаимосвязь между, например, дочерьми одного быка (чем больше общего между полусибсами), или, адекватно, чем больше влияние их общего отца.

Отношение σ_A^2 / σ_P^2 известно как «*коэффициент наследуемости в узком смысле*» - h^2 (очень важный в селекции животных генетический параметр популяции). Выборочная оценка h^2 по полусибсам:

$$\hat{h}^2 = 4 \times \hat{r}_w = \frac{4\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2} = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_P^2}.$$

$\hat{h}^2$ используют при расчете племенной ценности, например, по собственным показателям ($\hat{B}$ - оценка систематических эффектов среды):

$$EBV_i = \hat{h}^2 (y_i - \hat{\mu} - \hat{B}).$$

Точность EBV_i ($\hat{r}_{AA'}$) также рассчитывают из $\hat{h}^2$:

$$\hat{r}_{AA'} = \sqrt{\hat{h}^2};$$

$\hat{h}^2$ используют для прогноза ответа на селекции за поколение (ΔG):

$$\Delta G = i \times \hat{r}_{AA'} \times \hat{\sigma}_A = i \times \sqrt{\hat{h}^2} \times \sqrt{\hat{h}^2 \times \hat{\sigma}_y^2},$$

где i - стандартизированная интенсивность отбора.

Из последних четырех формул следует очень важный для практической селекции вывод - чем меньше $\hat{\sigma}_e^2$, тем: (1) больше генетическая изменчивость - $\hat{\sigma}_A^2$ и $\hat{h}^2$, (2) точнее оценка племенной ценности - $\hat{r}_{AA'}$ и (3) эффективнее селекционная работа - ΔG.

17.12. Предпосылки и преобразование данных

Основные предпосылки дисперсионного анализа:

- случайное, независимое и нормальное распределение компонентов ошибки (или анализируемой переменной);
- однородность варианс различных выборок;
- отсутствие корреляции между вариансами и средними различных выборок;
- слагаемость (аддитивность) главных эффектов.

Отклонение от нормального распределения анализируемой переменной и равенства варианс в ячейках (*если оно не чрезмерное*) - не сказывается существенно на результатах дисперсионного анализа при равном числе наблюдений в ячейках,

но может быть очень чувствительно при неравном их числе. Поэтому рекомендуется (1) планировать схему с равным числом наблюдений в ячейках, (2) если встречаются недостающие данные, то возмещать их средними значениями других наблюдений в ячейках. Искусственно введеные недостающие данные при подсчете числа степеней свободы не учитывают.

Аппарат дисперсионного анализа вполне применим, если распределение куполообразно и не резко асимметрично, а групповые варианты различаются в 1,5-2 раза. Допустимы и значительно большие отклонения, но при условии равенства и равномерности числа наблюдений в группах.

Если распределение значительно отличается от нормального, то необходимо преобразовать шкалу измерений так, чтобы новое распределение было близко к нормальному и имело стабильную вариансу.

В случае распределения Пуассона обычно используют преобразование $\sqrt{x}$ или, для малых значений *x*, преобразование $\sqrt{x+c}$, где оптимальным является c=0,386 (также достигается относительная независимость варианс от средних).

В случае биномиального распределения используют $2\arcsin\sqrt{p}$, где p - частота события (варианса стабилизируется для значений «p» от 0,05 до 0,95). Данные следует преобразовывать, если размах между отдельными значениями превышает 40%.

Для асимметричных, вытянутых вправо распределений, целесообразно преобразование $\lg x$.

Логарифмирование эффективно при наличии корреляции между стандартными отклонениями и средними (если все отношения дают близкие величины), а так же при неаддитивности главных эффектов. Если встречаются отрицательные или нулевые значения, то до лагорифмирования к ним прибавляют константу. Если встречаются значения меньше 1, то умножают на константу.

Все критерии проверки значимости и выделения существенных оценок проводят с преобразованными данными. Обратный перехода к исходным единицам делают лишь после получения оценок по преобразованным данным.

В заключение следует отметить, что данные преобразуют не для того, чтобы получить желательные исследователю результаты, а с целью обеспечения обоснованного анализа и получения правильных выводов.