

18. ДИСПЕРСИОННЫЙ АНАЛИЗ: ВЫБОР МОДЕЛИ

Цель дисперсионного анализа - разложить общую фенотипическую изменчивость признака на компоненты: изменчивость, вызванную воздействием учтенных в биометрической модели факторов, и изменчивость остаточную, возникшую под влиянием всех не учтенных факторов. В дисперсионном анализе различают биометрические модели *фиксированного*, *рандомизированного* и *смешанного* типов. Тип модели определяет подлежащие оценке параметры, влияет на интерпретацию результатов и репрезентативность (представительность) выводов. Поэтому перед обработкой данных с конкретной структурой необходимо определить тип биометрической модели [145, 146].

18.1. Модели фиксированного типа

Рассмотрим эксперимент по изучению влияния трех кормовых добавок (R1, R2, R3) на привес телят. Допустим, что эксперимент включает 24 животных. Три группы по 6 животных получают в рационе одну из кормовых добавок, а четвертая группа - контрольная (R4), не получает никакой добавки.

Для обработки этих данных биометрическая модель имеет следующий вид*:

$$y = \mu + R + e, \quad (1a)$$

где y - привес теленка; μ - общее среднее; R - эффект кормовой добавки; e - влияние неучтенных в модели факторов, остаток (ошибка).

Подобные эксперименты проводят для получения информации о достоверности или недостоверности влияния на продуктивность, воспроизводство или здоровье животных конкретных, представляющих интерес, факторов (тип кормления, способ содержания, метод лечения, система спаривания и т.п.). В рассматриваемом примере изучают три варианта кормовой добавки. Интерес представляет только сравнение R_1 , R_2 и R_3 друг с другом и с контролем (R_4). Внимание фокусируют только на вариантах данного

* Здесь и далее прописными буквами обозначены фиксированные, а строчными - рандомизированные факторы.

эксперимента. На них и ни на каких других. Поэтому эффекты вариантов (градаций) рационов являются фиксированными, а модель называют «модель с фиксированным фактором» или просто «фиксированной моделью», или «модель I».

Способ получения данных влияет на выводы. **Животные в каждом из вариантов опыта должны быть отобраны рандомизированно**. Тогда остатки (e) будут иметь среднее 0 и дисперсию σ_e^2 . Именно рандомизированная выборка для *каждого* варианта позволяет делать выводы о различиях между ними. Это является принципом фиксированной модели. Модель I используют для получения выводов об остаточной дисперсии. Она, среднее и эффекты факторов являются теми параметрами, которые подлежат оценке моделями данного типа.

Эксперимент можно расширить, повторив опыт в нескольких хозяйствах в сочетании с четырьмя вариантами кормовых добавок. Предположим, например, что использовались 4 вида кормовых добавок в 3-х хозяйствах: с высоким, средним и низким уровнями кормления. Биометрическая модель будет:

$$y = \mu + H + R + R \times H + e, \quad (16)$$

где H - эффект хозяйства (уровня кормления); $R \times H$ - эффект взаимодействия (совместного влияния) кормовой добавки и уровня кормления.

Как и эффект кормовой добавки, эффект уровня кормления (H) также является фиксированным, т.к. в этом эксперименте интерес сосредоточен только на трех конкретных хозяйствах. Таким образом, фиксированными факторами считают R , H и $R \times H$.

18.2. Модели рандомизированного типа

Рассмотрим эксперимент, в котором изучают материнский эффект у свиноматок. В качестве меры материнского эффекта используют вес приплода. Предположим, что эксперимент проводят на 4 свиноматках одной породы, случайно отобранных из стада. У каждой свиноматки по 6 поросят. Имея 24 поросенка, можно провести анализ изменчивости по следующей модели:

$$y = \mu + d + e, \quad (2a)$$

где y_{ij} - вес поросенка; d - эффект свиноматки (материнский эффект).

Изучаемый материнский эффект подвержен индивидуальной биологической изменчивости. Так как четыре свиноматки являются случайной выборкой из всего стада, то маловероятно, что интерес в эксперименте будет сосредоточен только на них.

Можно провести аналогию с предыдущим экспериментом - каждой свиноматке соответствует кормовая добавка. Но каждая кормовая добавка представляла интерес *сама по себе*. Мы не думали о ней, как о выборке из «совокупности кормовых добавок». В то время как каждая самка в данных о свиноматках является случайной выборкой (размером в единицу) из совокупности самок. Ничто важное не влияло на выбор какой-либо одной свиноматки по отношению к другой. И нас не интересуют различия между свиноматками, как это было в случае с кормовыми добавками. Наоборот, интерес в эксперименте со свиноматками заключается в том, чтобы определить - в какой степени материнский эффект варьирует во всей совокупности свиноматок (степень изменчивости материнского эффекта). Биометрическая модель 2а и предназначена для изучения этой изменчивости.

Выборка данных о свиноматках должна быть рандомизированной. Это предполагает, что остаточная вариация равна σ_e^2 . Таким образом, в отношении неучтенных факторов обе модели схожи. Различия заключаются в том, что в фиксированной модели планируют всегда иметь одни и те же варианты кормовых добавок. В то время как при использовании модели 2а в каждом случае необходимо брать *случайную* выборку свиноматок. Следовательно, материнский эффект (d) является рандомизированным фактором. Поэтому биометрическую модель 2а называют «моделью с рандомизированным фактором» или «рандомизированная модель», или «модель II».

Остаточные элементы в моделях I и II имеют одинаковые характеристики: средние равны нулю, дисперсии - σ_e^2 . Но если в модели I фактор R представлял эффект **конкретных** вариантов, то в модели II фактор d является **случайной** выборкой из совокупности, которая, как предполагают, имеет дисперсию σ_d^2 . Поэтому допускают, что эффекты d имеют среднее равное 0.

Модель II используют для получения выводов о σ_d^2 и σ_e^2 . Эти дисперсии и μ являются параметрами, которые подлежат оценке

моделью II. Из модели II следует, что $\sigma_y^2 = \sigma_d^2 + \sigma_e^2$. Поэтому σ_d^2 и σ_e^2 являются *компонентами* σ_y^2 . Исходя из этого, σ_d^2 и σ_e^2 называют «компонентами *фенотипической* *вариансы*».

Рандомизированная модель так же, как и фиксированная, может включать несколько факторов и их взаимодействие. Пример можно найти в молочном скотоводстве, когда в результате искусственного осеменения от быков можно получить потомство одновременно во многих стадах. Биометрическая модель будет иметь вид:

$$y = \mu + s + h + s \times h + e, \quad (2b)$$

где y - продуктивность первотелки; s - влияние на продуктивность первотелки генотипа отца; h - эффект стада (хозяйства); $s \times h$ - эффект взаимодействия бык \times стадо («генотип \times среда»).

В этой модели все эффекты рандомизированные. Быки считаются рандомизированной выборкой из некоторой совокупности быков. Стада считаются рандомизированной выборкой из совокупности стад. Поэтому эффекты взаимодействия также предполагаются рандомизированными. И, конечно, влияния неучтенных факторов являются случайными. Допускают, что все эффекты не коррелируют друг с другом и имеют varianсы σ_s^2 , σ_h^2 , σ_{sh}^2 и σ_e^2 . Они являются компонентами фенотипической varianсы модели 2b:

$$\sigma_y^2 = \sigma_s^2 + \sigma_h^2 + \sigma_{sh}^2 + \sigma_e^2.$$

Варianсы и μ - это те параметры, которые подлежат оцениванию биометрическими моделями рандомизированного типа.

Иногда из-за небольшого размера выборки и/или сильной несбалансированности данных, и/или недостаточной адекватности модели данным, и/или неадекватности принятых допущений **некоторые оценки варianс могут иметь отрицательные значения**. В этом случае рекомендуют:

- приравнять нулю отрицательные оценки варianс, остальные оценки оставить без изменения, или
- приравнять нулю отрицательные оценки варianс, остальные оценки пересчитать, или

- отбросить отрицательные варианты как незначимые, оставшиеся оценить заново, или
- интерпретировать исследуемую модель как неправильную, сделать анализ по новой модели, или
- интерпретировать результат как недостаток данных, или
- использовать другой метод оценивания.

18.3. Модели смешанного типа

Вернемся к эксперименту с кормовыми добавками и хозяйствами (модель 1b). Но сейчас допустим, что вместо 3-х хозяйств с разным уровнем продуктивности, в эксперименте участвуют животные из 20 *случайно* выбранных хозяйств. В этом случае биометрическая модель будет иметь вид:

$$y = \mu + h + R + r \times h + e, \quad (3)$$

где h - эффект хозяйства; R - эффект кормовой добавки; $r \times h$ - эффект взаимодействия между кормовой добавкой и хозяйством.

Как и ранее, R является фиксированным фактором. Однако фактор h , представляющий эффекты 20 хозяйств, является рандомизированным. Эти 20 хозяйств были случайно отобраны для эксперимента из всей совокупности хозяйств. Следовательно, модель содержит как фиксированные (R), так и рандомизированные (h) факторы. Такую модель называют «моделью со смешанными факторами» или «смешанной моделью», или «модель III».

Биометрическая модель 3 включает эффект взаимодействия между вариантом кормовой добавки и хозяйством - $r \times h$. Так как допускают, что h - рандомизированный фактор, то логично считать это взаимодействие также рандомизированным. Таким образом, модель 3 включает R - как фиксированный эффект, h и $r \times h$ - как рандомизированные эффекты со средними равными 0 и вариансами σ_h^2 и σ_{rh}^2 . Поэтому параметрами, которые подлежат оцениванию, являются μ , эффекты фиксированного фактора R и варианты рандомизированных факторов: σ_h^2 , σ_{rh}^2 и σ_e^2 .

18.4. Фиксированная или рандомизированная модель?

Фиксированные уровни фактора - это уровни, выбор которых делают **сознательно**. Обычно это уровни, которые представляют для исследователя интерес. Между фиксированными уровнями предполагаются постоянные (систематические) отличия - эффекты. В этом случае можно высказать строго обоснованные суждения об особенностях каждого уровня. Выводы распространяются только на эти уровни.

Эффекты могут рассматриваться как фиксированные, когда число уровней относительно небольшое и ограничивается этим числом при повторном осуществлении выборки. Таким образом, если уровни (градации) факторов точно установлены (три рациона кормления, четыре сезона отела) или характеризуют определенное состояние (самцы, самки), или если оценивают влияние отдельных градаций фактора из всех возможных (четыре конкретных стада из двадцати), то такие факторы являются фиксированными. Например, при исследовании молочной продуктивности испытывают четыре новых рациона, заведомо различающихся своими свойствами. Необходимо сравнить их по эффективности и выбрать наилучший. В данном исследовании интерес представляют только эти четыре рациона и никакие другие. Поэтому выводы должны распространяться только на эти рационы.

Рандомизированные уровни выбирают случайным образом из большого или бесконечного числа однотипных уровней. Различия между такими уровнями должны быть только случайными. При этом предполагают, что число возможных уровней, из которых сделана выборка, достаточно велико. Поэтому выводы могут распространяться на всю популяцию, из которой произведена выборка. Эффекты животных часто считают рандомизированными, т.к. при повторении эксперимента в выборке будут другие животные, случайно отобранные из популяции.

Фиксированные или рандомизированные. Считать ли уровни фиксированными или рандомизированными во многом зависит от контекста и целей исследования. В одном контексте эффекты могут быть фиксированными, в другом - рандомизированными.

Предположим, что свиноматок обслуживают три оператора, и что влияние оператора необходимо учесть в биометрической модели. Если задача эксперимента заключается в оценке *различий между тремя операторами* с целью выявления лучшего, то три оператора являются единственными операторами, представляющими интерес. Они не должны рассматриваться как случайная выборка из совокупности операторов. Поэтому эффекты операторов необходимо считать фиксированными.

С другой стороны, за проведением эксперимента кто-то должен следить. Применительно к самому эксперименту нас мало интересует, кто обслуживает животных. Любого оператора можно рассматривать как случайную выборку из всех возможных операторов. В этом эксперименте требуется оценить лишь *вклад операторов в общую изменчивость* данных о свиноматках. Поэтому эффекты операторов в модели необходимо рассматривать как рандомизированные. Тогда оценке подлежит дисперсия, обусловленная операторами.

Ответы на следующие ниже вопросы могут помочь в принятии решения о том, каким считать фактор - фиксированным или рандомизированным:

1. Число градаций (уровней, эффектов) фактора?

Небольшое - фиксированный.

Большое - возможно, рандомизированный.

2. Имеют ли градации фактора повторности?

Да - фиксированный.

Нет - рандомизированный.

3. Может ли в принципе число градаций фактора быть бесконечно большим?

Да - возможно, рандомизированный.

Нет - возможно, фиксированный.

4. Случайный ли выбор градаций фактора?

Да - рандомизированный.

Нет - фиксированный.

5. Будут ли выводы распространяться на популяцию или на выборку, по которой были получены оценки?

Да - рандомизированный.

Нет - фиксированный.

В заключение следует отметить, что допущение о рандомизированном характере эффектов не подразумевает допущение об их нормальном распределении. Большинство процедур оценивания компонентов общей дисперсии не требуют нормальности распределения. Но при исследовании свойств оценок параметров, нормальность распределения обычно предполагается.