Использование множественных измерений в таксономических задачах

Р. А. Фишер,

доктор наук,

член Лондонского королевского общества

Аннотация

Эта работа написана для иллюстрации практического численного примера из области растительной таксономии, где понятие дискриминантной функции, по-видимому, может оказать непосредственную пользу.

1 Дискриминантные функции

Когда две или более популяции измерены по нескольким признакам

$$x_1, \ldots, x_s,$$

особый интерес представляют определённые линейные функции этих измерений, с помощью которых популяции лучше всего различаются. По предложению автора уже было использовано это обстоятельство в краниометрии: (a) мистером Э. С. Мартином, который применил принцип к половым различиям в измерениях нижней челюсти, и (b) мисс Милдред Барнард, которая показала, как из серии датированных наборов получить особую комбинацию черепных измерений, наиболее явно демонстрирующую прогрессивную или секулярную тенденцию. В настоящей работе применение того же принципа будет проиллюстрировано на таксономической задаче; также будут обсуждены некоторые вопросы, связанные с точностью применяемых процедур.

2 Арифметическая процедура

В Таблице I приведены измерения цветков по пятидесяти растениям каждого из двух видов *Iris setosa* и *I. versicolor*, растущих вместе в одной колонии и измеренных доктором Э. Андерсоном, которому я обязан за предоставленные данные. Приведены четыре измерения цветка.

Сначала рассмотрим вопрос: какую линейную функцию четырёх измерений

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$$

следует выбрать, чтобы максимизировать отношение разности между средними значениями видов к стандартному отклонению внутри видов? Наблюдаемые средние значения и их различия приведены в Таблице II. Разности можно обозначить через d_p, где $p = 1$, 2, 3 или 4 для четырёх измерений.

Суммы квадратов и произведений отклонений от средних значений каждого вида показаны в Таблице III. Поскольку использовались по пятьдесят растений каждого вида, эти суммы содержат 98 степеней свободы. Эти суммы квадратов или произведений можно обозначить через S_{pq}, где p и q независимо принимают значения 1, 2, 3 и 4.

Тогда для любой линейной функции X измерений, определённой выше, разность между средними значениями X у двух видов равна

$$D = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 + \lambda_4 d_4$$

в то время как дисперсия X внутри видов пропорциональна

$$S = \sum_{p=1}^{4} \sum_{q=1}^{4} \lambda_p \lambda_q S_{pq}$$

.

Особая линейная функция, которая наилучшим образом различает два вида, будет той, для которой отношение D^2/S максимально, при независимом изменении четырёх коэффициентов λ_1, λ_2, λ_3 и λ_4. Это даёт для каждого λ

$$\frac{D}{S^2}\left\{2S\frac{\partial D}{\partial \lambda} - D\frac{\partial S}{\partial \lambda}\right\} = 0,$$

или

Таблица I

Iris setosa				*Iris versicolor*				*Iris virginica*			
Чашелистик		Лепесток		Чашелистик		Лепесток		Чашелистик		Лепесток	
Длина	Ширина	Длина	Ширина	Длина	Ширина	Длина	Ширина	Длина	Ширина	Длина	Ширина
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.1	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.1	1.5	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Таблица II: Наблюдаемые средние значения для двух видов и их разность, см

	versicolor	*setosa*	Разность $(V - S)$
Длина чашелистика (x_1)	5.936	5.006	0.930
Ширина чашелистика (x_2)	2.770	3.428	-0.658
Длина лепестка (x_3)	4.260	1.462	2.798
Ширина лепестка (x_4)	1.326	0.246	1.080

Таблица III: Суммы квадратов и произведений четырёх измерений внутри видов, см^2

	Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка
Длина чашелистика	19.1434	9.0356	9.7634	3.2394
Ширина чашелистика	9.0356	11.8658	4.6232	2.4746
Длина лепестка	9.7634	4.6232	12.2978	3.8794
Ширина лепестка	3.2394	2.4746	3.8794	2.4604

$$\frac{1}{2}\frac{\partial S}{\partial \lambda} = \frac{S}{D}\frac{\partial D}{\partial \lambda},$$

где можно заметить, что S/D является постоянным множителем для четырёх неизвестных коэффициентов. Следовательно, требуемые коэффициенты пропорциональны решениям системы уравнений

$$\left.\begin{aligned} S_{11}\lambda_1 + S_{12}\lambda_2 + S_{13}\lambda_3 + S_{14}\lambda_4 &= d_1, \\ S_{21}\lambda_1 + S_{22}\lambda_2 + S_{23}\lambda_3 + S_{24}\lambda_4 &= d_2, \\ S_{31}\lambda_1 + S_{32}\lambda_2 + S_{33}\lambda_3 + S_{34}\lambda_4 &= d_3, \\ S_{41}\lambda_1 + S_{42}\lambda_2 + S_{43}\lambda_3 + S_{44}\lambda_4 &= d_4. \end{aligned}\right\} \tag{1}$$

Если, в свою очередь, подставить единицу для каждой из разностей и ноль для остальных, полученные решения составляют матрицу множителей, обратную матрице S; численно мы находим:

Таблица IV: Матрица множителей, обратная суммам квадратов и произведений внутри видов (см$^{-2}$).

	Длина чашелистика	Ширина чашелистика	Длина лепестка	Ширина лепестка
Длина чашелистика	0.1187161	-0.0668666	-0.0816158	0.0396350
Ширина чашелистика	-0.0668666	0.1452736	0.0334101	-0.1107529
Длина лепестка	-0.0816158	0.0334101	0.2193614	-0.2720206
Ширина лепестка	0.0396350	-0.1107529	-0.2720206	0.8945506

Эти значения можно обозначить как s_{pq} для p и q от 1 до 4.

Умножив столбцы матрицы из Таблицы IV на наблюдаемые разности, получаем решения уравнения (1) в виде

$$\lambda_1 = -0.0311511, \lambda_2 = -0.1839075, \lambda_3 = +0.2221044, \lambda_4 = +0.3147370,$$

так что, если принять коэффициент длины чашелистика за единицу, требуемая составная величина равна

$$X = x_1 + 5.9037x_2 - 7.1299x_3 - 10.1036x_4.$$

Если в этом выражении подставить наблюдаемые значения у растений *setosa*, среднее, вычисленное по данным из Таблицы I, будет

$$5.006 + (3.428)(5.9037) - (1.462)(7.1299) - (0.246)(10.1036) = 12.3345 \text{ см};$$

для *versicolor*, наоборот, имеем

$$5.936 + (2.770)(5.9037) - (4.260)(7.1299) - (1.326)(10.1036) = -21.4815 \text{ см}.$$

Разность средних значений составной величины таким образом равна 33.816 см.

Явность метрических признаков двух видов теперь можно оценить, сравнив эту разность средних со стандартной ошибкой. Используя значения из Таблицы III с коэффициентами нашей составной величины, получаем

$$\begin{aligned}
19.1431 + (9.0356)(5.9037) - (9.7634)(7.1299) - (3.2394)(10.1036) &= -29.8508, \\
9.0356 + (11.8658)(5.9037) - (4.6232)(7.1299) - (2.4746)(10.1036) &= 21.1224, \\
9.7634 + (4.6232)(5.9037) - (12.2978)(7.1299) - (3.8794)(10.1036) &= -89.8206, \\
3.2394 + (2.4746)(5.9037) - (3.8794)(7.1299) - (2.4604)(10.1036) &= -34.6699,
\end{aligned}$$

и, наконец,

$$-29.8508+(21.1224)(5.9037)+(89.8206)(7.1299)+(34.6699)(10.1036) = 1085.5522.$$

Среднюю дисперсию двух видов по составной величине можно оценить, разделив это значение (1085.5522) на 95; дисперсию разности между двумя средними по пятидесяти растениям каждого вида — разделив снова на 25. Для одного растения дисперсия равна 11.4269, так что средняя разность 33.816 см между парой растений разных видов имеет стандартное отклонение 4.781 см. Для средних по пятидесяти растениям та же средняя разность имеет стандартную ошибку 0.6761 см, что составляет примерно одну пятидесятую от её значения.

3 Интерпретация

Отношение разности средних выбранной составной величины к её стандартной ошибке для отдельных растений также представляет интерес в связи с вероятностью ошибочной классификации, если природа вида оценивалась бы исключительно по измерениям. По причинам, которые будут обсуждены далее, дисперсию одного растения мы оценим, разделив 1085.5522 на 95, получив 11.4269 см^2 для дисперсии и 3.3804 см для стандартного отклонения. Предположим, что растение классифицировано неверно, если его отклонение в правильном направлении превышает половину разности, 33.816 см, между видами; отношение к стандартному отклонению, как оценено, равно 5.0018.

Таблица нормального распределения (*Статистические методы*, Таблица II) показывает, что отношение 4.89164 превышается пять раз на миллион, а 5.32672 — только один раз на два миллиона испытаний. По логарифмической интерполяции частота, соответствующая отношению 5.0018, составляет примерно 2.79 на миллион. Если дисперсии двух видов не равны, эта частота несколько переоценивается данным методом, поскольку следует делить специфическую разность пропорционально двум стандартным отклонениям, а при постоянной сумме дисперсий сумма стандартных отклонений наибольшая, когда они равны. Следовательно, можно сразу заключить, что если измерения распределены почти нормально, вероятность ошибочной классификации, используя только составную величину, составляет менее трёх на миллион.

То же отношение интересно и с другой стороны. Если выбранная составная величина X анализируется с точки зрения её вариации внутри и между видами, сумма квадратов между видами должна быть равна $25D^2$. Численно мы имеем, следовательно,

Таблица V: Анализ дисперсии выбранной составной величины X, между и внутри видов

	Степени свободы	Сумма квадратов
Между видами	4	28588.05
Внутри видов	95	1085.55
Итого	99	29673.60

Из общей суммы только 3.6583% приходится на внутривидовую вариацию, а 96.3417% — на межвидовую. Составная величина выбрана так, чтобы максимизировать последнюю долю. Поскольку, кроме специфических средних, мы использовали три настраиваемых коэффициента, вариация внутри видов содержит лишь 95 степеней свободы.

При составлении вариативной величины X мы умножили исходные значения λ на -32.1018, чтобы придать измерению длины чашелистика коэффициент, равный единице. Если бы мы использовали исходные значения, анализ из Таблицы V выглядел бы следующим образом:

Таблица VI: Анализ дисперсии исходной составной величины X, между и внутри видов

	Степени свободы	Сумма квадратов	
Между видами	4	27.74160	$= 25D^2$
Внутри видов	95	1.05341	$= D = S$
Итого	99	28.79501	$D(1 + 25D)$

При умножении уравнений (1) на λ_1, λ_2, λ_3 и λ_4 и последующем сложении оказывается, что $S = \sum \lambda d = D$, то есть специфическая разность в исходной составной величине X. Долю суммы квадратов внутри видов (3.6 %) можно было бы найти просто как $1/(1 + 25D)$.

4 Аналогия частичной регрессии

Анализ Таблицы VI наводит на интересную аналогию. Если каждому растению присвоить значение вариативной величины y, одинаковое для

всех представителей одного вида, анализ дисперсии y, между частями, объясняемыми линейной регрессией по измерениям $x_1, \ldots, x_4$, и остаточная вариация после подгонки такой регрессии будут идентичны Таблице VI, если y задать соответствующие равные и противоположные значения для двух видов.

В общем случае, при различном числе представителей двух видов, n_1 и n_2, если значения y, присвоенные растениям, равны

$$\frac{n_2}{n_1+n_2} \quad \text{и} \quad \frac{-n_1}{n_1+n_2},$$

отличаясь на единицу, правые части уравнений для коэффициентов регрессии, соответствующих уравнению (1), будут

$$\frac{n_1 n_2}{n_1+n_2} d_p,$$

где d_p — разность средних двух видов по любому из измерений. Типичный коэффициент левой части будет

$$S_{pq} + \frac{n_1 n_2}{n_1+n_2} d_p d_q.$$

Перенеся дополнительные дроби в правую часть, мы получим уравнения, идентичные (1), за исключением того, что правые части теперь равны

$$\frac{n_1 n_2}{n_1+n_2} d_p \,(1 - \sum \lambda' d),$$

где λ' обозначает решение новых уравнений; следовательно

$$\lambda' = \frac{n_1 n_2}{n_1+n_2} (1 - \sum \lambda' d)\, \lambda,$$

умножаем эти уравнения на d и суммируем, так что

$$\sum \lambda' d = \frac{n_1 n_2}{n_1+n_2} \sum \lambda d \,(1 - \sum \lambda' d),$$

или

$$(1 - \sum \lambda' d) \left(1 + \frac{n_1 n_2}{n_1+n_2} \sum \lambda d\right) = 1,$$

и, следовательно, в нашем примере

$$1 - \sum \lambda' d = \frac{1}{1 + 25D}.$$

Анализ дисперсии y будет, таким образом,

Таблица VII: Анализ дисперсии вариативной величины y, определяемой исключительно видом

	Степени свободы	Сумма квадратов	
Регрессия	4	24.0854	$25^2 D/(1 + 25D)$
Остаток	95	0.9146	$25/(1 + 25D)$
Итого	99	25.0000	

Общая сумма $S(y^2)$, как видно, в общем случае равна $\frac{n_1 n_2}{n_1 + n_2}$; часть, приписываемая регрессии, равна

$$\frac{n_1 n_2}{n_1 + n_2} \sum \lambda' d = \frac{25^2 D}{1 + 25D}.$$

В этом способе представления соответствующее распределение степеней свободы очевидно.

Множественная корреляция y с измерениями $x_1, \ldots, x_4$ задаётся формулой

$$R^2 = \frac{25D}{1 + 25D}.$$

5 Проверка значимости

Теперь становится ясно, каким образом специфическую разность можно проверить на значимость, учитывая, что вариативная величина выбрана так, чтобы максимизировать различимость видов. Регрессия y по четырём измерениям имеет 4 степени свободы, а остаточная вариация — 95; значение z, вычисленное по суммам квадратов в любой из Таблиц V, VI или VII, равно 3.2183 или

$$\frac{1}{2}(\log\ 95 - \log\ 4 + \log\ 25 + \log\ D),$$

что является весьма значимым значением для числа использованных степеней свободы.

6 Применение к теории аллополиплоидии

Теперь можно рассмотреть одно из расширений этой процедуры, доступных, когда выборки взяты более чем из двух популяций. Выборка третьего вида, приведённая в Таблице I, *Iris virginica*, отличается от двух других выборок тем, что не была взята из той же естественной колонии, что и они — обстоятельство, которое может значительно исказить как средние значения, так и их вариативность. Интерес представляет её соотношение с *I. setosa* и *I. versicolor*, поскольку Randoph (1934) установил [1], а Anderson подтвердил [2], что в то время как *I. setosa* является «диплоидным» видом с 38 хромосомами, *I. virginica* — «тетраплоидным» с 70 хромосомами, а *I. versicolor*, который является промежуточным по трём измерениям, хотя не по ширине чашелистика, — гекса-плоидным. Он предположил интересную возможность, что *I. versicolor* является полиплоидным гибридом двух других видов. Мы, следовательно, рассмотрим, принимает ли среднее значение для *I. versicolor* промежуточное значение при использовании линейной составной величины из четырёх измерений, наиболее подходящей для различения трёх таких видов, и если да, то отличается ли оно в два раза больше от *I. setosa*, чем от *I. virginica*, как можно было бы ожидать, если эффекты генов просто аддитивны, в гибриде между диплоидным и тетраплоидным видами.

Если третье значение находится на двух третях пути от одного значения к другому, три отклонения от их общего среднего должны находиться в соотношении $4 : 1 : -5$. Чтобы получить значения, соответствующие разностям между двумя видами, мы можем, следовательно, формировать линейные составные величины их средних измерений с использованием этих числовых коэффициентов. Результаты приведены в Таблице VIII, где, например, значение 7.258 см для длины чашелистика равно четырём средним длинам чашелистика для *I. virginica*, плюс одному среднему для *I. versicolor* минус пять средних значений для *I. setosa*.

Таблица VIII

Средние значения	S_{pq}			
Iris virginica. Пятьдесят растений				
6.588	19.8128	4.5944	14.8612	2.4056
2.974	4.5944	5.0962	3.4976	2.3338
5.552	14.8612	3.4976	14.9258	2.3924
2.026	2.4056	2.3338	2.3924	3.6962
Iris versicolor. Пятьдесят растений				
5.936	13.0552	4.1740	8.9620	2.7332
2.770	4.1740	4.8250	4.0500	2.0190
4.260	8.9620	4.0500	10.8200	3.5820
1.326	2.7332	2.0190	3.5820	1.9182
Iris setosa. Пятьдесят растений				
5.006	6.0882	4.8816	0.8014	0.5062
3.428	4.8616	7.0408	0.5732	0.4556
1.462	0.8014	0.5732	1.4778	0.2974
0.246	0.5062	0.4556	0.2974	0.5442
$4vi + ve - 5se$				
7.258	482.2650	199.2244	266.7762	53.8778
-2.474	199.2244	262.3842	74.3416	50.7498
19.158	266.7762	74.3416	286.6618	49.2954
8.200	53.8778	50.7498	49.2954	74.6604

Поскольку значения сумм квадратов и произведений отклонений от средних внутри каждого из трёх видов несколько различаются, мы можем составить соответствующую матрицу для выбранной линейной составной величины, умножив значения для *I. virginica* на 16, для *I. versicolor* на 1, а для *I. setosa* на 25, и сложив значения для трёх видов, как показано в Таблице VIII. Полученные таким образом значения будут соответствовать матрице сумм квадратов и произведений внутри видов, если выборки брались только из двух популяций.

Используя строки матрицы как коэффициенты четырёх неизвестных в уравнении с нашей выбранной составной величиной средних измерений, например

$$482.2650\lambda_1 + 199.2244\lambda_2 + 266.7762\lambda_3 + 53.8778\lambda_4 = 7.258,$$

Мы находим решения, которые, если умножить на 100, будут равны

Коэффициент длины чашелистика	−3.308998
ширины чашелистика	−2.759132
длины лепестка	8.866048
ширины лепестка	9.392551

определяя тем самым требуемую составную величину.

Теперь легко найти средние значения и дисперсии этой составной величины для трёх видов. Они приведены в таблице ниже (Таблица IX):

Таблица IX

	Среднее	Сумма квадратов	Средний квадрат	Стандартное отклонение
$I.virginica$	38.24827	923.7958	18.8530	4.342
$I.versicolor$	22.93888	873.5119	17.8268	4.222
$I.setosa$	-10.75042	292.8958	5.9775	2.444

Из этой таблицы видно, что, в то время как разность между *I. setosa* и *I. versicolor*, 33.69 наших единиц, настолько велика по сравнению со стандартными отклонениями, что значительного перекрытия значений не может происходить, разность между *I. virginica* и *I. versicolor*, 15.31 единиц, меньше чем в четыре раза стандартное отклонение каждого вида.

Тем не менее, разности, похоже, удивительно близки к отношению 2:1. По сравнению с этой нормой, *I. virginica* кажется оказавшей слегка преобладающее влияние. Отклонение от ожидания, однако, невелико, и у нас есть материал для проведения хотя бы приблизительной проверки значимости.

Если бы разности между средними точно соответствовали соотношению 2:1, то линейная функция, образованная сложением средних с коэффициентами в соотношении $2 : -3 : 1$, была бы равна нулю. На самом деле она равна 3.07052. Дисперсия выборки этой составной величины находится умножением дисперсий трёх видов на 4, 9 и 1, последующим сложением и делением на 50, поскольку каждое среднее основано на пятидесяти растениях. Это даёт 4.8365 для дисперсии и 2.199 для стандартной ошибки. Таким образом, по этому тесту расхождение 3.071,

безусловно, не является значимым, хотя оно несколько превышает стандартную ошибку.

В теории тест на значимость не является полностью точным, так как при оценке дисперсии выборки каждого вида мы делили сумму квадратов отклонений от среднего на 49, как если бы эти отклонения имели всего 147 степеней свободы. На самом деле три степени свободы были поглощены при корректировке коэффициентов линейной составной величины так, чтобы максимально чётко различать виды. Если бы мы делили на 48 вместо 49, стандартная ошибка была бы слегка увеличена до значения 2.231, что не повлияло бы на интерпретацию данных. Однако такое изменение, безусловно, было бы излишней коррекцией, так как именно дисперсии крайних видов *I. virginica* и *I. setosa* наиболее сокращаются при выборе составной величины, в то время как дисперсия *I. versicolor* вносит основную часть ошибки выборки в тесте на значимость.

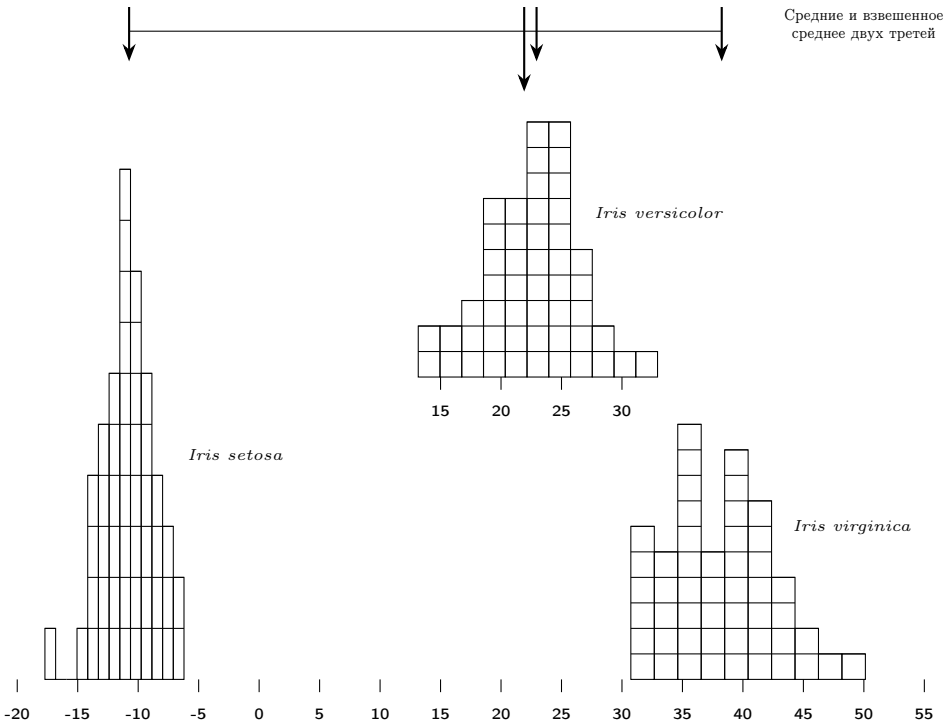

Рис. 1: Гистограммы частот дискриминирующей линейной функции для трёх видов *Iris*.

На диаграмме, рисунок 1, показаны фактические распределения составной величины, принятой для отдельных особей трёх измеряемых видов. Как было предсказано выше, заметно, что наблюдается некоторое перекрытие распределений *I. virginica* и *I. versicolor*, так что точная диагностика этих двух видов не может основываться исключительно на этих четырёх измерениях одного цветка, взятого с растения, растущего в природе. Тем не менее, в культуре возможно, что одни лишь измерения обеспечат более полное различение видов.

Русские названия видов ирисов

- *Iris setosa* — Ирис щетинистый
- *Iris versicolor* — Ирис разноцветный
- *Iris virginica* — Ирис виргинский

Список литературы

[1] Randolph, L. F. (1934). "Chromosome numbers in native American and introduced species and cultivated varieties of Iris." *Bull. Amer. Iris Soc.* **52**, 61–66.

[2] Anderson, Edgar (1935). "The irises of the Gaspe Peninsula." *Bull. Amer. Iris Soc.* **59**, 2–5.

[3] Anderson, Edgar (1936). "The species problem in *Iris*." *Ann. Mo. Bot. Gdn.* (in press).