Занятие 10 . Изучение межпопуляционных различий

Изолированные популяции могут накапливать адаптивные или случайные морфологические различия. Поиск маркёров изолированных группировок может осуществляться методом дискриминантного анализа. Исходным материалом данной работы являются морфометрические данные полученные от самок трёх группировок мух журчалок, живущих в долинах горных рек Таланда, Сигикта и Сутору (левые притоки Горина, Комсомольский заповедник), пойманных в один и тот же полевой сезон. Задачей анализа является решение вопросов:

- есть ли морфометрические различия между представителями трёх популяций?
- есть ли способ определения принадлежности мух к конкретной группировке?

Работа осуществляется в программе STATISTICA (с иллюстрациями интерфейса STATISTICA.7)

Дискриминантный анализ

С чем работает дискриминантный анализ. Задача метода в некотором смысле обратна задаче кластерного анализа: имеются объекты с определенными признаками; необходимо, зная группировки объектов, найти комбинации признаков, по которым можно сказать, к какой группировке объект относится; предполагается, что зная эти признаки, *каждый* объект из генеральной совокупности можно отнести к определенной группировке с достаточно высокой вероятностью. В отличие от других методов анализа, вопрос о том, различаются ли группы по данному признаку, второстепенен - важно не то, различается ли амебный менингит и клещевой энцефалит по средней температуре первой недели заболевания, а по каким признакам для *каждого* больного можно поставить надежный диагноз. Однако, если группировок несколько, приходится допустить, что некоторые из них имеют больше различий, чем другие и эти различия желательно задать количественно.

Ограничения дискриминантного анализа. Метод разработан при допущении, что все признаки распределены нормально и в случае корреляции связь между ними линейна. Однако на практике часто оказывается возможным использовать в дискриминантном анализе дискретные или даже качественные признаки. Более того, одной из побочных задач дискриминантного анализа может быть оптимальная оцифровка упорядоченных качественных признаков[1]. Метод хуже работает, если признаки сильно скоррелированы

[1] Упорядоченный или ранжированный качественный признак подразумевает наличие связей между переменными типа "больше-меньше". Например, среди группы характеристик

друг с другом. Поэтому избыточные признаки лучше сразу определить и удалить из дискриминантной модели.

Как работает дискриминантный анализ. Исходные данные вводятся в виде таблицы, где строка определяет объект, а колонка - признак.

Предварительной или побочной задачей может быть **определение достоверности различий между группировками**. Интуитивно ожидается возможность того, что несколько признаков в сумме могут дать надежное определение, хотя каждый из них, взятый по отдельности, недостаточен для диагноза. С другой стороны, несколько кажущихся важными признаков могут быть так тесно скоррелированы между собой, что их количество ничего не решает.

Обе проблемы снимаются, если мы переходим к рассмотрению группировок в пространстве Махаланобиса, которое является пространством многомерного нормального распределения. Здесь, наряду с определением расстояний между группировками можно получить и вероятности нуль-гипотезы - утверждения, что группировки неразличимы (истинное расстояние равно нулю, а наблюдаемое является следствием неполноты выборки).

Собственно дискриминантный анализ осуществляется двумя основными методами:

• классический метод - вычисление *линейных дискриминантных функций*. Каждая группировка представляется в виде облака точек в многомерном пространстве (количество измерений равно количеству признаков), и это облако определяется линией регрессии, заданной дискриминантной функцией. Каждый объект является точкой этого пространства. Точка должна быть отнесена к той группировке, к регрессионной линии которой она ближе всего расположена. Конечной целью анализа является вычисление не регрессионных (дискриминантных), а *классификационных* функций. Каждая группировка имеет свою классификационную функцию - набор коэффициентов, на которые умножаются значения соответствующих признаков. Произведения суммируются и объект должен быть отнесен к той группировке, для которой эта сумма больше.

Точнее, эти функции имеют вид

$$S_i = x_1 \times w_{1i} + x_2 \times w_{2i} + \ldots x_m \times w_{mi} + c_i$$

где x_k - это значение k-ого признака объекта, w_{ki} - коэффициент i-ой группировки для k-ого признака, c_i - свободный член i-ой группировки.

Вычислительные сложности здесь кажущиеся - расчеты легко организовать в любой электронной таблице.

"турок", "немец", "поляк", "еврей" таких связей нет, а среди "да", "скорее да, чем нет", "скорее нет, чем да", "нет" - есть.

• *канонический дискриминантный анализ*. Классификационные функции вычисляются методом канонической корреляции. Количество функций определяется количеством значимых корней канонической корреляции; как правило, их много меньше, чем исходных признаков. Они задают новое пространство, в котором определяется "центр тяжести" каждой группировки - *центроид*. Объект, определенный в этом пространстве как точка, относится к той группировке, к центроиду которой он расположен ближе.

Поскольку канонических корней как правило, не более трех, малопонятный неспециалисту "табличный" результат анализа допускает простую и очевидную графическую интерпретацию. Это большое преимущество канонического дискриминантного анализа перед линейным.

Так же как в факторном анализе, корни канонического дискриминантного анализа допускают возможность их интерпретации как некоторой "скрытой сущности" (для этого разыскиваются признаки, наиболее тесно скоррелированные с корнями), однако обычно это не делается.

Проверка эффективности дискриминации апостериорно (т.е. "задним числом") учитывает количество ошибочных классификаций. Это производится как напрямую, с учетом ошибок для каждой группировки, так и косвенно, с вычислением *Λ*-критерия Уилка. Этот показатель меняется от нуля до единицы, причем 0 - абсолютно точная классификация, 1 - абсолютно ошибочная.

Отбор признаков осуществляется несколькими способами:

• методом пошаговой регрессии

• оценкой толерантности признака; толерантность - степень нескоррелированности признака со всеми остальными, величина изменчивости признака, которую нельзя оценить по другим признакам. Если толерантность признака близка к нулю, от него лучше избавиться

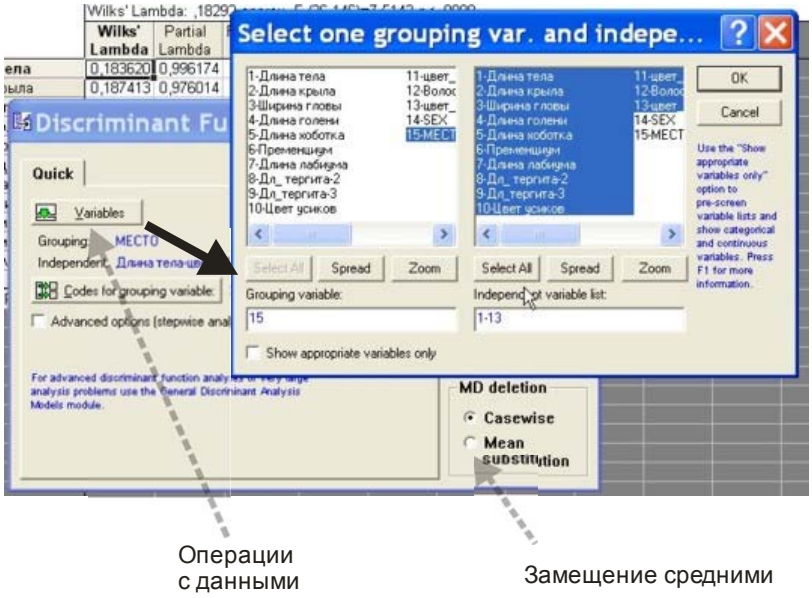
• вычислением частного значения *Λ*-критерия Уилкса - показателя того, насколько признак в одиночку способен выполнять классифицирующую функцию

• определением вероятности нуль-гипотезы, предполагающей, что при удалении признака точность классификации не изменится.

***Задание 1*. Каноническая дискриминация - разбор примера**. Загрузите файл ollfem.sta в модуль *Discriminant analysis (Statistics – Multivariate Exploratory Techniques - Discriminant analysis)*. Описание файла: у самок мух сем. журчалок пойманных в 3 местообитаниях - долинах речек Таланда, Сигикта, Сутору почти поровну в каждом месте, изучены 13 признаков - 9 количественных (длина крыла, ширина головы и т. д.) и 4 ранжированных качественных - с задачей: определить, есть ли

межпопуляционная изменчивость мух по этим показателям, достаточная для определения местообитания мухи.

Ввод данных. В исходном окне *Discrimimant Function Analysis (Анализ дискриминантной функции)* в оконце, открываемом кнопкой *Variables (Переменные)* в левой панели *Grouping variable (Переменная, задающая группировки)* выделите МЕСТО, в правой панели (*Independent variable list - Список независимых переменных*) - все 13 признаков. В оконце *Missing data (Пропущенные переменные)* можно поставить режим *Mean substitution - Замещение средними значениями* (некоторые ячейки таблицы данных пусты; хотя заполнять их средними значениями не вполне честно, часто это лучше, чем выбрасывать целые строки в режиме *Casewise - Удаление [строк с пропущенными данными]*).



Оценка признаков. После нажатия клавиши Ок появится окно *Discriminant function Analysis Results* (*Аналитические результаты дискриминантной функции*) на закладке *Quick* (*Краткий*) с единственной активной кнопкой *Summary: Variables in Model* (*Итог: Переменные модели*). Отметим, что дискриминантные функции будут работать достаточно надёжно: *Wilks' Lambda* ≈ *0,18* – это хороший показатель.

Щелкните по кнопке *Summary: Variables in Model.* В нашем случае окажется, что из тринадцати использованных признаков только пять выделены красным цветом. Это значит, что от оставшихся восьми можно избавиться, они не информативны.

Classification Matrix (OLLFEM.STA)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	Таланда p=,34091	Сигикта p=,31818	Сутору p=,34091
Таланда	56,66667	17	10	3
Сигикта	96,42857	0	27	1
Сутору	93,33334	1	1	28
Total	81,81818	18	38	32

Оценка эффективности определения по дискриминантным функциям. Перейдите на закладку *Classification* и щёлкните кнопку *Classification matrix* (*Классификационная таблица*). В появившейся таблице представлены результаты работы классификационной функции. Заголовки в белом поле можно переписать так: Rows: Observed classifications – Ряды: исходная классификация, Columns: Predicted classifications – Колонки: предсказанная классификация.

Итак, машина построила какие-то правила работы с нашими признаками (какие именно – разберём ниже) и по этим правилам попыталась определить принадлежность мухи к какой-либо популяции, используя только признаки, т.е., как бы не ведая, откуда дровишки. И вот что получилось: Мухи из долины Таланды правильно идентифицированы в ≈ 57% случаев – 17 как таландинские, 10 как сигиктовские и 3 как суторинские. Мухи из Сигикты правильно определены в ≈ 96% случаев: 0 как таландинские, 27 как сигикитнские и 1 как суторинская и т.д. Общий процент успеха – Total/Percent correct ≈ 82%. Если отбросить 8 «избыточных» признаков и просчитать всё снова, то результат несколько ухудшится: Wilks' Lambda ≈ 0,22152, Total/Percent correct ≈ 80%. Упростив наблюдения, потеряем в точности классификации 2%.

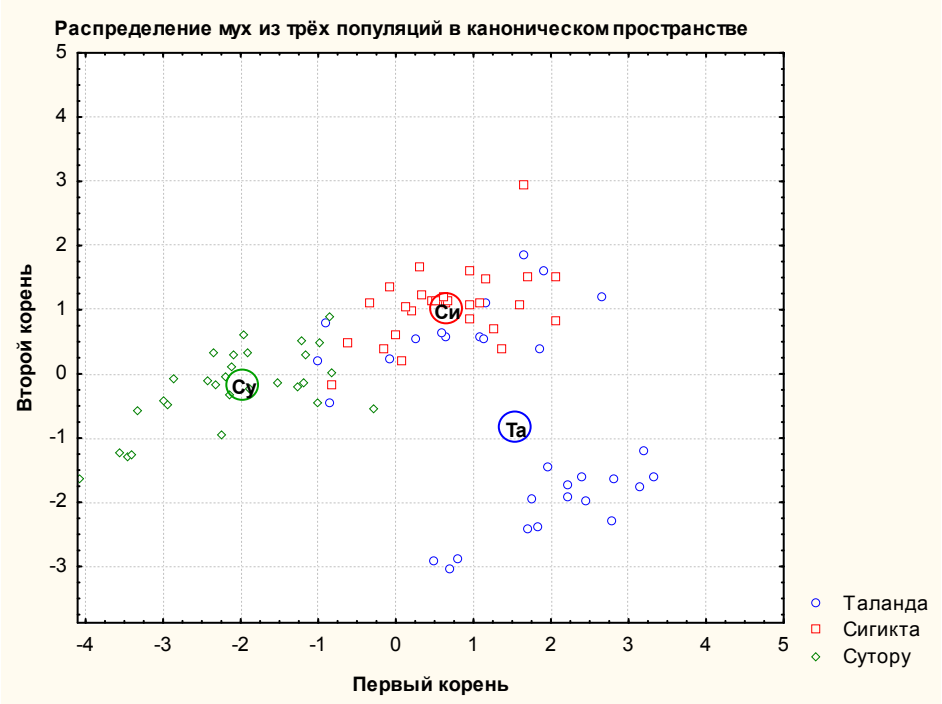
В некоторых случаях, например при диагностике болезней, ошибка классификации может стоить дорого, однако степень надежности классификации для каждого объекта можно определить по таблице, вызываемой кнопкой *Posterior probabilities (Апостериорные вероятности)*:

Posterior Probabilities (OLLFEM_1.STA)
Incorrect classifications are marked with *

Case	Observed Classif.	Таланда p=,34091	Сигикта p=,31818	Сутору p=,34091
10	Таланда	0,994693	0,005305	0,000002
11	Таланда	0,979047	0,005370	0,015582
12	Таланда	0,994999	0,004826	0,000175
* 13	Таланда	0,357731	0,638580	0,003689
* 14	Таланда	0,224828	0,758018	0,017155
* 15	Таланда	0,041001	0,222465	0,736534
* 16	Таланда	0,031593	0,454675	0,513733
17	Таланда	0,993860	0,006139	0,000001
18	Таланда	0,997294	0,002700	0,000006
* 19	Таланда	0,227169	0,628928	0,143902
* 20	Таланда	0,168722	0,829048	0,002230

Дискриминантные функции. Дискриминантный анализ есть поиск сложных математических функций, которые позволяют путём хитроумных расчётов определить, к какому классу относится объект, имеющий некоторый набор признаков. Однако обычно вычислять их не требуется. Из двух разновидностей дискриминатных функций – классических и канонических – остановимся на последних.

Щёлкните кнопку *Perform canonical analysis* (*Осуществить канонический анализ*). В появившемся окне *Canonical analysis* выберите закладку *Canonical scores* и щёлкните кнопку *Scatterplot of canonical scores* (*График канонических ценок*). Полученный график почти полностью разъясняет ситуацию. В доработанном виде он может выглядеть так:



Ясно, что мухи из популяций ТАЛАНДА морфологически неоднородны, их различие учитывает второй корень. Наоборот, мух из популяции СУТОРУ можно определить достаточно надежно.

По непонятной причине составители программы не сочли необходимым вывод на график центроидов. Координаты центроидов можно получить в таблице, выводимой кнопкой *Means of Canonical Variables (Средние канонических переменных)* из закладки Advanced (Расширенный) и вручную ввести (нарисовать) центроиды в диаграмму, например так, как это сделано на приведенном выше графике.

Теперь идея метода понятна чисто интуитивно: точка относится к той группе, к центроиду которой она ближе.

Как вычисляются координаты точек? В данном примере координатная сеть образована двумя каноническими корнями (их вообще может быть и больше). Каждый корень даёт набор множителей для исходных признаков. Значения признаков перемножаем на эти коэффициенты, складываем, и получаем координату точки по данному корню.

Нужно ли их считать? Сама по себе работа не так трудоёмка, если использовать формулы, введённые в электронные таблицы. Однако и её выполнять необязательно.

Классификация новых объектов по дискриминантным функциям. Разумеется, прикладное значение дискриминантного анализа состоит в том, чтобы на основании экспериментальных данных можно было бы составить правила, по которым можно было бы классифицировать новый материал. Щелкнув правой клавишей мыши по последней строке, получите выпадающее меню, в котором выбором опций *Modify Case(s) - Add (Модифицировать строки - Добавить)* вызовите оконце *Add Cases (Добавление строк),* где в позицию *Number to Cases to Add (Число добавляемых строк)* поставьте любую цифру больше 0 (достаточно 1). В таблице данных появятся пустые строки. Введите в них значения признаков для некоторой теоретической мухи, например 10, 9, 3, 2.5, 1.75, .75, 1.25, 3.75, 3.50, 2, 3, 1, но оставьте пустой ячейку в колонке МЕСТО. Сбросьте старые установки, повторите вычисления и просмотрите таблицу апостериорных вероятностей. "Научившись" на исходных данных, STATISTICA определит вероятности классификаций и для новой строки с неизвестным местообитанием.

Стоит ли исследовать корни? Координаты точек выводятся кнопкой *Coefficients for canonical variables (Коэффициенты для канонических переменных),* расположенной в окне *Canonical analysis* . При щелчке на кнопке появляются две таблицы - с "сырыми" (*row*) и стандартизированными (*standartized*) коэффициентами. Пользоваться "сырыми" следует почти так же, как и линейными дискриминантными функциями - значение каждого признака исследуемого объекта умножается на соответствующий коэффициент, произведения складываются, прибавляется константа (*constant*) и получается координата данной точки по оси соответствующего корня. Удобнее всего создать соответствующую формулу в электронных таблицах. Тогда при вводе признаков автоматически будут подсчитываться координаты объекта и его расстояние до центроидов. Последняя строка таблицы - *Cum. prop. (Накопленный процент)* показывает долю общей изменчивости (точнее, дисперсии) объясненной данным корнем.

Стандартизированные данные имеют то преимущество, что они сравнимы друг с другом; их можно использовать при анализе значимости признаков. Это значит, что чем больше коэффициент признака по абсолютной величине, тем больше признак важен для дискриминации (если, конечно, он не скоррелирован с другим значимым признаком).

Выше отмечалось, что мухи из долины Таланды фактически разорваны на две группировки, причём этот разрыв определяется только вторым корнем (ось *Root 2*). Какие признаки из стандартизованных коэффициентов второго корня наиболее значимы? «Цвет фемура» и «Волоски на лице». Следовательно, имеет смысл проверить, не по ним ли происходит разрыв? Для этого дос-

таточно выбросить их из анализа и построить новый график точек на канонических корнях.

Variable	Standardized Coefficien for Canonical Variables	
	Root 1	**Root 2**
Длина тела	-0,071460	0,098895
Длина крыла	-0,259615	-0,102282
Ширина гловы	0,375583	0,049833
Длина голени	-0,409360	0,117515
Длина хоботка	-0,458550	-0,117160
Пременциум	-0,024052	0,321644
Длина лабиума	-0,297798	-0,074720
Дл_ тергита-2	0,255343	-0,168005
Дл_тергита-3	-0,193629	0,062019
Цвет усиков	0,439307	-0,144044
цвет_фемура	-0,653331	0,489469
Волоски_на_лице	0,377419	0,832839
цвет_МТР	0,205463	-0,205428
Eigenval	2,434462	0,615786
Cum.Prop	0,798119	1,000000

Задание 2. Проведите самостоятельный анализ только по количественным данным того же файла.