


# Занятие 9

ОСНОВЫ МНОГОМЕРНЫХ  
МЕТОДОВ АНАЛИЗА.  
MANOVA.

Дискриминантный анализ.

# Общие принципы многомерного анализа

**Многомерные данные:** несколько переменных регистрируются для каждого объекта в выборке (особи, образца, ...)



Много **независимых** переменных

- ✓ Многофакторная ANOVA
- ✓ Множественная регрессия

Много **ЗАВИСИМЫХ** переменных (или переменных, которые нельзя разделить на зависимые и независимые) –  
✓ **multivariate analyses**

Перейдём к ситуации, когда проверяется влияние **одной** или **нескольких независимых** переменных на **НЕСКОЛЬКО ЗАВИСИМЫХ** переменных.

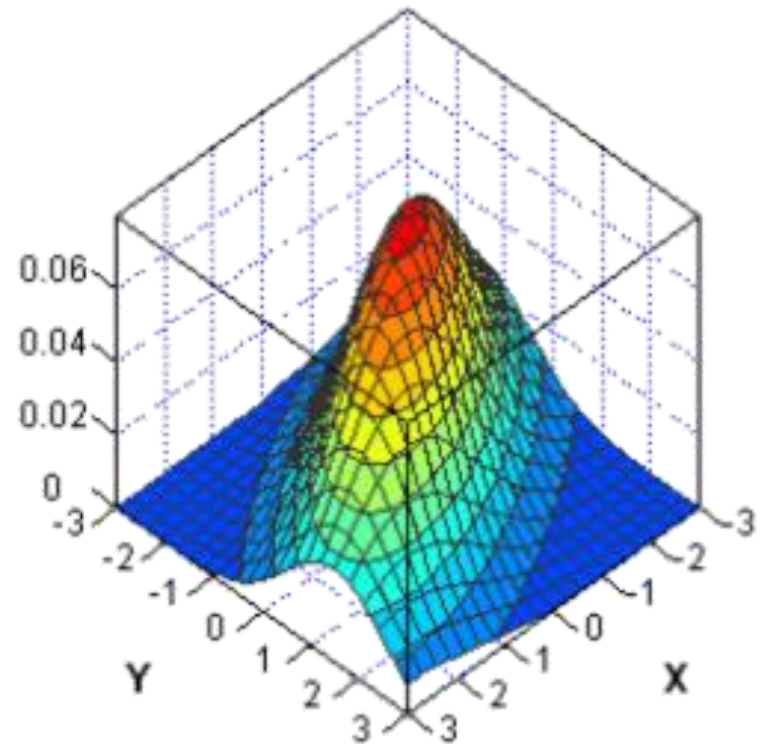
Наши данные:  $n$  объектов, для каждого измерено  $p$  переменных.

## Описание многомерных данных

### 1. Распределение многомерных данных – многомерное.

При тестировании гипотез в многомерном анализе требуется **многомерное нормальное распределение** (это значит, все переменные и их линейные комбинации распределены нормально).

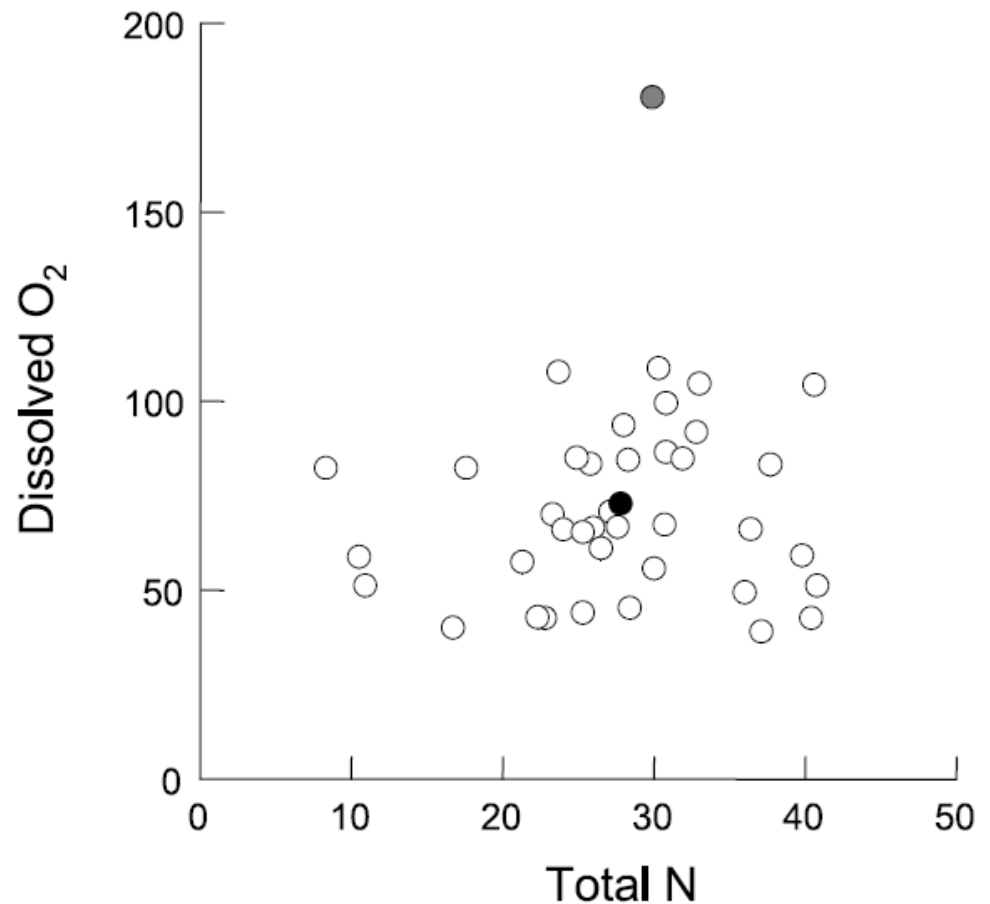
Чем больше отклонение от многомерного нормального распределения, тем больше будет неточности в оценке параметров (коэффициентов и пр.).



**2. Показатель «середины»** распределения: для одной переменной – среднее значение.

Для многомерного распределения – **ЦЕНТРОИД**. Точка, координаты которой – средние значения для каждой переменной.

Для каждого объекта можно посчитать его «расстояние» до центроида (дистанция Махаланобиса).



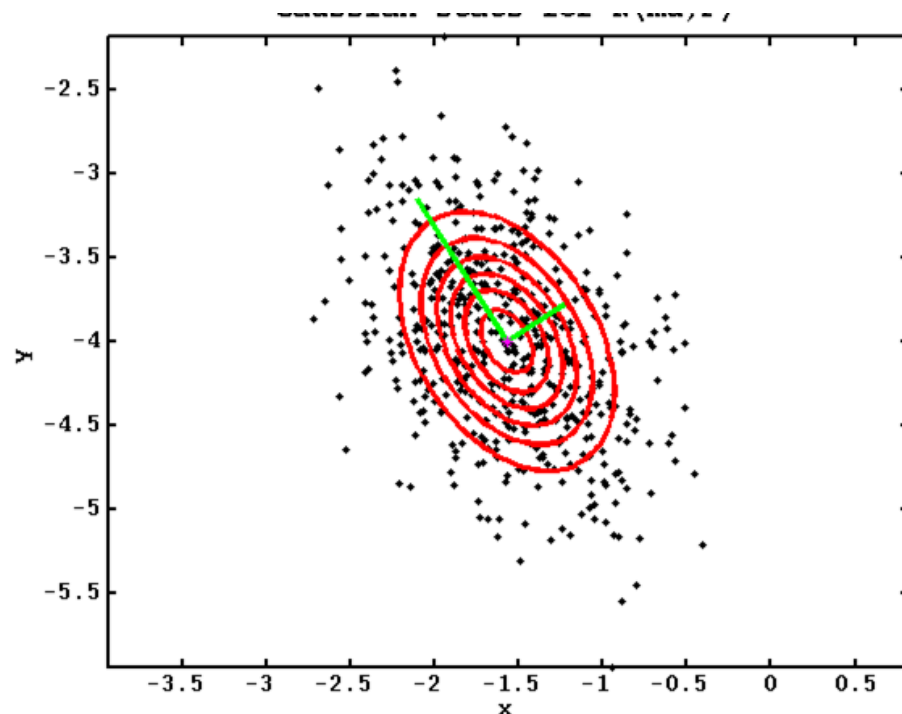
## Описание многомерных данных

3. Показатели **разброса**: в одномерном распределении – сумма квадратов отклонений (SS), дисперсия, стандартное отклонение.

Трудность в том, в многомерных данных **два** источника изменчивости:

- ✓ изменчивость **внутри самих переменных**;
- ✓ изменчивость, обусловленная **взаимным влиянием переменных**.

*Что же делать?*



## Описание многомерных данных

Изменчивость в многомерных данных представляется сложно – в виде таблицы (**матрицы**).

Немного о **матрицах** (матрицы – основа многомерного анализа!):

- ✓ Это прямоугольные **таблицы**, которые состоят из **чисел** (элементов).
- ✓ В матрицах есть **строки** и **столбцы** (нумеруются слева-направо, сверху-вниз)

$$A = (3 \quad -10 \quad 0.5, \quad 0.1) \quad A = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix}$$

$$A = \begin{pmatrix} 4 & 5 & -1 & 2 \\ 6 & 0 & 2 & -3 \end{pmatrix} \quad A = \begin{pmatrix} 1 & 3 & -7 \\ 0 & 2 & 8 \\ -5 & 1 & 0 \end{pmatrix}$$

**Рис. 1**

Это матрица	Это строка матрицы	Это столбец матрицы																																				
<table><tr><td>3</td><td>8</td><td>47</td></tr><tr><td>20</td><td>5</td><td>79</td></tr><tr><td>3</td><td>53</td><td>0</td></tr><tr><td>6</td><td>22</td><td>1</td></tr></table>	3	8	47	20	5	79	3	53	0	6	22	1	<table><tr><td>3</td><td>8</td><td>47</td></tr><tr><td>20</td><td>5</td><td>79</td></tr><tr><td>3</td><td>53</td><td>0</td></tr><tr><td>6</td><td>22</td><td>1</td></tr></table>	3	8	47	20	5	79	3	53	0	6	22	1	<table><tr><td>3</td><td>8</td><td>47</td></tr><tr><td>20</td><td>5</td><td>79</td></tr><tr><td>3</td><td>53</td><td>0</td></tr><tr><td>6</td><td>22</td><td>1</td></tr></table>	3	8	47	20	5	79	3	53	0	6	22	1
3	8	47																																				
20	5	79																																				
3	53	0																																				
6	22	1																																				
3	8	47																																				
20	5	79																																				
3	53	0																																				
6	22	1																																				
3	8	47																																				
20	5	79																																				
3	53	0																																				
6	22	1																																				

## Описание многомерных данных

- ✓  $m \times n$  матрица – **прямоугольная**;  $n \times n$  – **квадратная**;
- ✓ у каждого **элемента** есть **номер** строки и столбца, в которых он стоит;
- ✓ у квадратных матриц есть **диагональ**
- ✓ с матрицами и их строками/столбцами можно производить всякие **действия**: менять строки/столбцы местами; умножать на число; прибавлять число; складывать и умножать матрицы; переворачивать относительно диагонали (транспонировать).

Первая строка стала первым столбцом

$$A = \begin{pmatrix} -1 & 2 & 4 & 0 & 7 \\ 3 & -5 & 24 & 9 & -3 \\ -10 & -8 & -2 & -4 & 11 \end{pmatrix}$$

$$A^T = \begin{pmatrix} -1 & 3 & -10 \\ 2 & -5 & -8 \\ 4 & 24 & -2 \\ 0 & 9 & -4 \\ 7 & -3 & 11 \end{pmatrix}$$

Вторая строка стала вторым столбцом

Третья строка стала третьим столбцом

$$\begin{pmatrix} 2 & -2 & 9 & 1 \\ 5 & 9 & 8 & 0 \\ 1 & 0 & 4 & -7 \\ -4 & -9 & 5 & 6 \end{pmatrix}$$

главная  
диагональ

## Описание многомерных данных

Мы на свою **таблицу с данными** можем посмотреть, как на **матрицу**: в ней есть столбцы, строки, она прямоугольная.

Clevenger & Waltho изучали, сколько раз и как (на велосипеде-верхом-пешком) люди переходят дорогу в заповеднике на разных 11 переходах.

Underpass	Raw		
	Bicycle	Horse	Foot
1	0	6	7
2	5	3	45
3	6	6	14
4	21	5	20
5	189	42	34
6	8	138	77
7	462	186	129
8	19	12	80
9	595	58	241
10	1	10	10
11	0	10	29



$$\begin{bmatrix} y_{11} & y_{12} & \dots & y_{1p} \\ y_{21} & y_{22} & \dots & y_{2p} \\ \dots & \dots & y_{ij} & \dots \\ y_{n1} & y_{n2} & \dots & y_{np} \end{bmatrix}$$

## Описание многомерных данных

Чтобы описать **изменчивость** многомерных данных, нам понадобится **матрица**, так как нам надо показать и изменчивость внутри переменных, и их взаимодействие – **каждой** переменной **с каждой**.

Матрица будет **квадратной**,  $p \times p$ , где  $p$  – число переменных, и **симметричной** относительно диагонали.

Во-первых, матрица sums-of-squares-and-cross-products (**SSCP**) (неудобна, т.к. сильно зависит от абсолютных значений):

$$\begin{bmatrix} \sum_{i=1}^n (y_{i1} - \bar{y}_1)^2 & \sum_{i=1}^n (y_{i2} - \bar{y}_2)(y_{i1} - \bar{y}_1) & \dots & \sum_{i=1}^n (y_{ip} - \bar{y}_p)(y_{i1} - \bar{y}_1) \\ \sum_{i=1}^n (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2) & \sum_{i=1}^n (y_{i2} - \bar{y}_2)^2 & \dots & \sum_{i=1}^n (y_{ip} - \bar{y}_p)(y_{i2} - \bar{y}_2) \\ \dots & \dots & \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2 & \dots \\ \sum_{i=1}^n (y_{i1} - \bar{y}_1)(y_{ip} - \bar{y}_p) & \sum_{i=1}^n (y_{i2} - \bar{y}_2)(y_{ip} - \bar{y}_p) & \dots & \sum_{i=1}^n (y_{ip} - \bar{y}_p)^2 \end{bmatrix}$$

На главной диагонали – **SS**, остальные – произведения отклонений в парах переменных

## Описание многомерных данных

Основные матрицы – **матрица ковариаций**...

$$\begin{bmatrix} s_1^2 & s_{12}^2 & \dots & s_{p1}^2 \\ s_{12}^2 & s_2^2 & \dots & s_{p2}^2 \\ \dots & \dots & s_j^2 & \dots \\ s_{1p}^2 & s_{2p}^2 & \dots & s_p^2 \end{bmatrix}$$

- ✓ на главной диагонали стоят **дисперсии** для каждой переменной – показатели разброса **внутри** переменных;
- ✓ остальные элементы – **ковариации** (covariances, C) между переменными – показатели **взаимосвязи между** переменными.

p переменных, n объектов

$$s_1^2 = \frac{\sum_{i=1}^n (Y_{i1} - \bar{Y}_1)^2}{n-1}$$

$$s_{12}^2 = \frac{\sum_{i=1}^n (Y_{i1} - \bar{Y}_1)^2 (Y_{i2} - \bar{Y}_2)^2}{n-1}$$

Дисперсия 1-й переменной      Ковариация 1-й и 2-й переменных

## Описание многомерных данных

	Bicycle	Horse	Foot
Bicycle	44 906.018		
Horse	7336.382	3862.018	
Foot	13 084.709	2205.191	4903.655



## Матрица ковариаций

Если нужно выразить общий **разброс** одним числом, используют:

1. **Сумму дисперсий** от всех переменных (сумма элементов диагонали; «след» матрицы, trace);
2. Сумму перемноженных особым образом элементов разных строк и столбцов – **определитель** матрицы.

## Описание многомерных данных

...и **матрица корреляций** (correlation matrix, R).

$$\begin{bmatrix} 1 & r_{21} & \dots & r_{p1} \\ r_{12} & 1 & \dots & r_{p2} \\ \dots & \dots & 1 & \dots \\ r_{1p} & r_{2p} & \dots & 1 \end{bmatrix}$$

Мы уже встречались с ней в регрессиях и корреляциях.

На главной диагонали – **единицы**, всё остальное – **коэффициенты корреляции** в парах переменных.

Она получится, если в предыдущей матрице (ковариаций) каждый элемент поделить на его стандартное отклонение.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y}$$



## Многомерный анализ

Основная роль этих матриц – путём простых преобразований получить на основе исходных переменных **НОВЫЕ ПЕРЕМЕННЫЕ**.

Новые переменные – **ЛИНЕЙНЫЕ КОМБИНАЦИИ** исходных, такие, что общая **изменчивость** по-новому распределяется между ними.

Т.е., для каждого объекта будет своё значение новой переменной (для  $i$ -го (от 1 до  $n$ ) объекта,  $p$  исходных переменных можно рассчитать значение новой  $k$ -той переменной как):

$$z_{ik} = c_1 y_{i1} + c_2 y_{i2} + \dots + c_j y_{ij} + \dots + c_p y_{ip}$$

Здесь  $y$  – значения исходных переменных для данного объекта,  $c$  – коэффициенты, показывающие величину вклада данной исходной переменной в новую переменную. В некоторых моделях добавляют ещё константу - intercept

## Многомерный анализ

Получение новых переменных из линейной комбинации всех исходных —

**основная техника** и ядро всех многомерных методов.

**В разных методах** их называют:

- ✓ дискриминантные функции (discriminant functions);
- ✓ канонические функции (canonical functions);
- ✓ вариаты (variates);
- ✓ главные компоненты (principal components);
- ✓ факторы (factors);
- ✓ корни (roots).

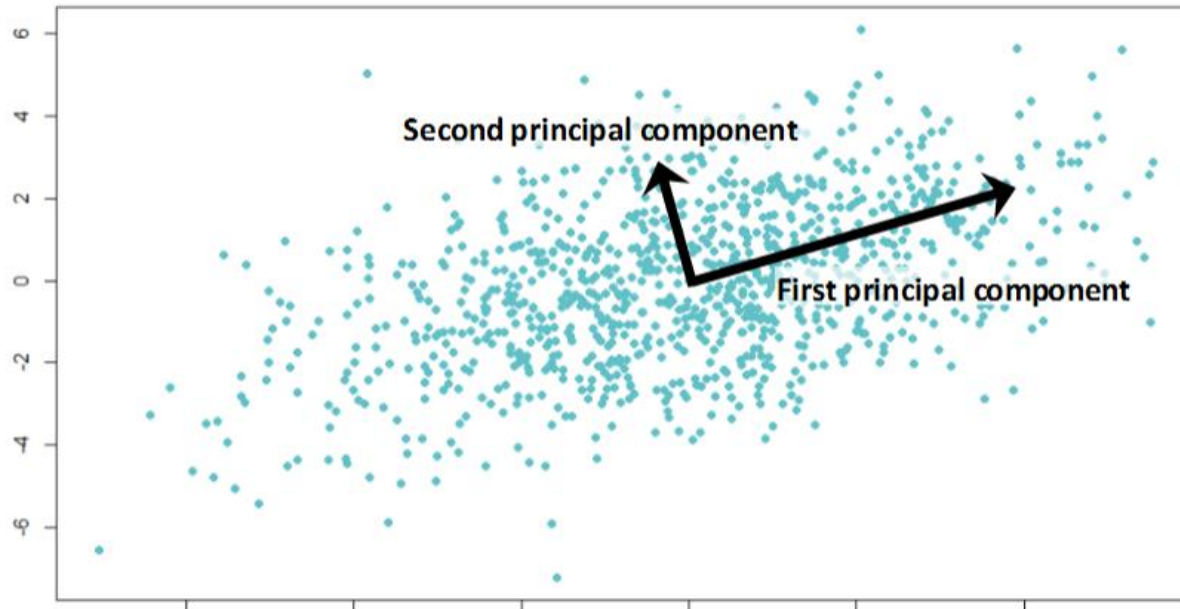


Это уравнение очень похоже на уравнение линейной регрессии, как будто мы сами делаем новую зависимую переменную.

## Многомерный анализ

### Свойства новых переменных

- ✓ На **первую** приходится **максимум изменчивости** исходных переменных, на вторую – максимум оставшейся изменчивости, и.т.д.
- ✓ таким образом, большая часть общей дисперсии оказывается в нескольких первых;
- ✓ они **не коррелируют** друг с другом;
- ✓ их  **$p$**  штук (т.е., столько, сколько исходных переменных).



# Многомерный анализ

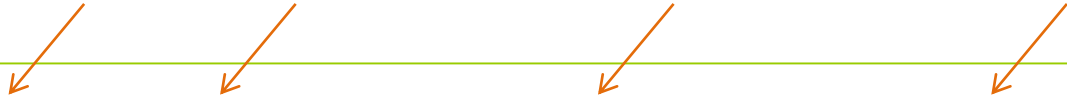
*У новых переменных есть:*

✓ Собственное значение ( $\lambda$ ) = **eigenvalue**, показывает, какая доля общей изменчивости приходится на переменную. Это популяционные параметры, у них есть выборочные оценки –  $l$ .

Их сумма = **сумме дисперсий** (если мы их строим на основе матрицы ковариаций), или числу исходных переменных (для матрицы корреляций).

Eigenvalue = characteristic roots, latent roots

✓ Собственный вектор = **eigenvector**, список коэффициентов при исходных переменных.


$$z_{ik} = c_1 y_{i1} + c_2 y_{i2} + \dots + c_j y_{ij} + \dots + c_p y_{ip}$$

## Многомерный анализ

Эти новые переменные (линейные комбинации) получаются с помощью простых действий над матрицами: «разложение» матрицы (ковариаций или корреляций  $p \times p$ ) разом даёт матрицу с **eigenvectors** и матрицу с **eigenvalues**.

Собственные значения для новых переменных

50 075.681	0	0
0	2592.350	0
0	0	1003.660



Eigenvector	1	2	3
Eigenvalue	50 075.681	2592.350	1003.660
Percentage of total variance	93.300	4.830	1.870

	1	2	3
Bicycle	0.945	0.160	0.284
Horse	0.164	-0.986	0.011
Foot	0.282	0.036	-0.959

Коэффициенты для новых переменных (столбец = eigenvector)

## Многомерный анализ

Теперь можно для каждого объекта (перехода) посчитать значения новых переменных = компонент. И, например, использовать в дальнейшем анализе.

Мы рассмотрели способ получения компонент (и их значений для объектов) из матриц ковариаций или корреляций ( $p \times p$ ).

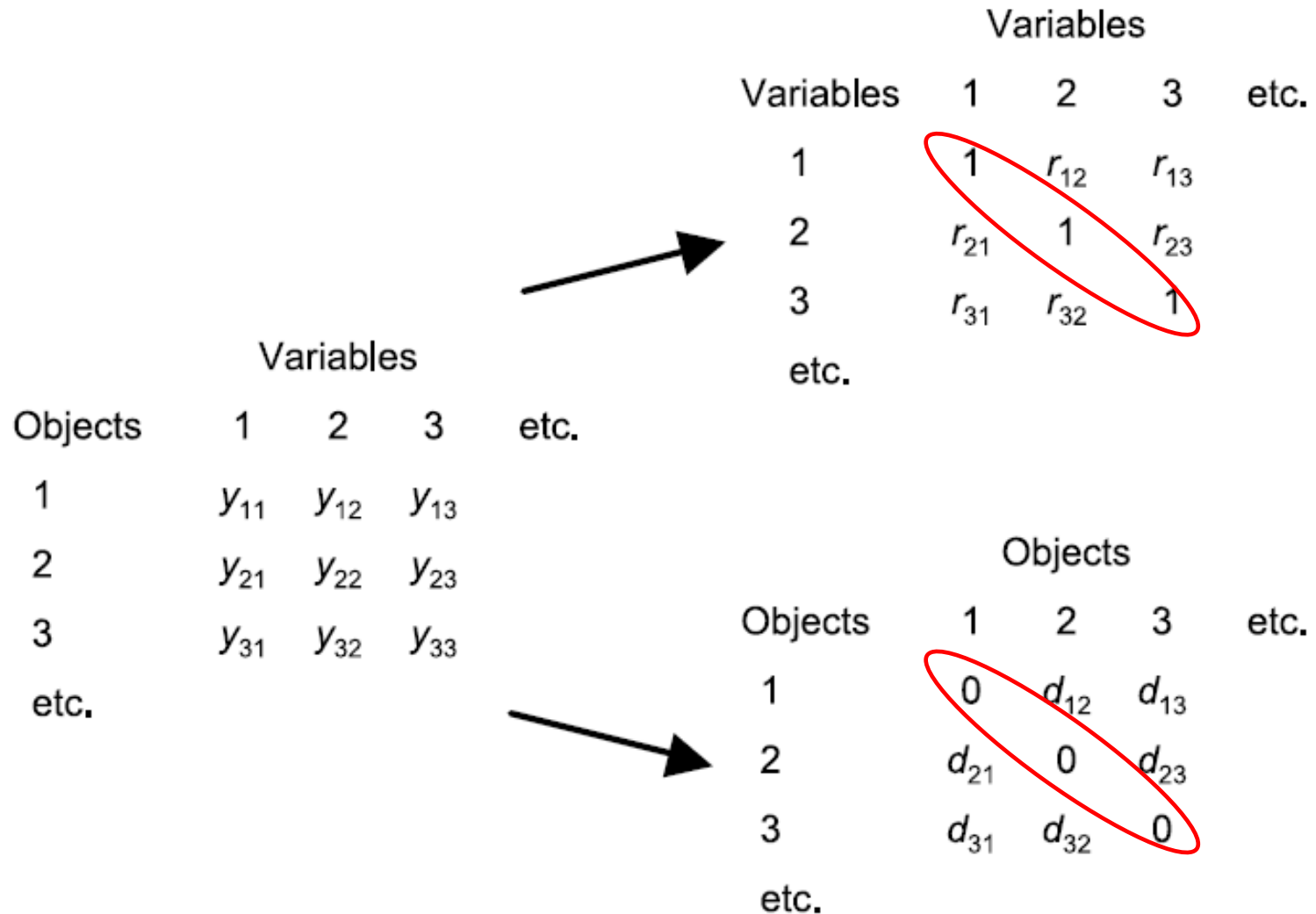
– **R-mode analysis.**

Есть другой способ: построить матрицу «корреляций» = «дистанций» между объектами ( $n \times n$ ) в пространстве исходных переменных, и из этой матрицы (тоже путём «разложения» матрицы) рассчитать значения новых компонент (они будут другие), и затем найти eigenvectors - **Q-mode analysis.**

Разные пути используются в разных типах многомерного анализа, но вообще-то они алгебраически связаны.

## Многомерный анализ

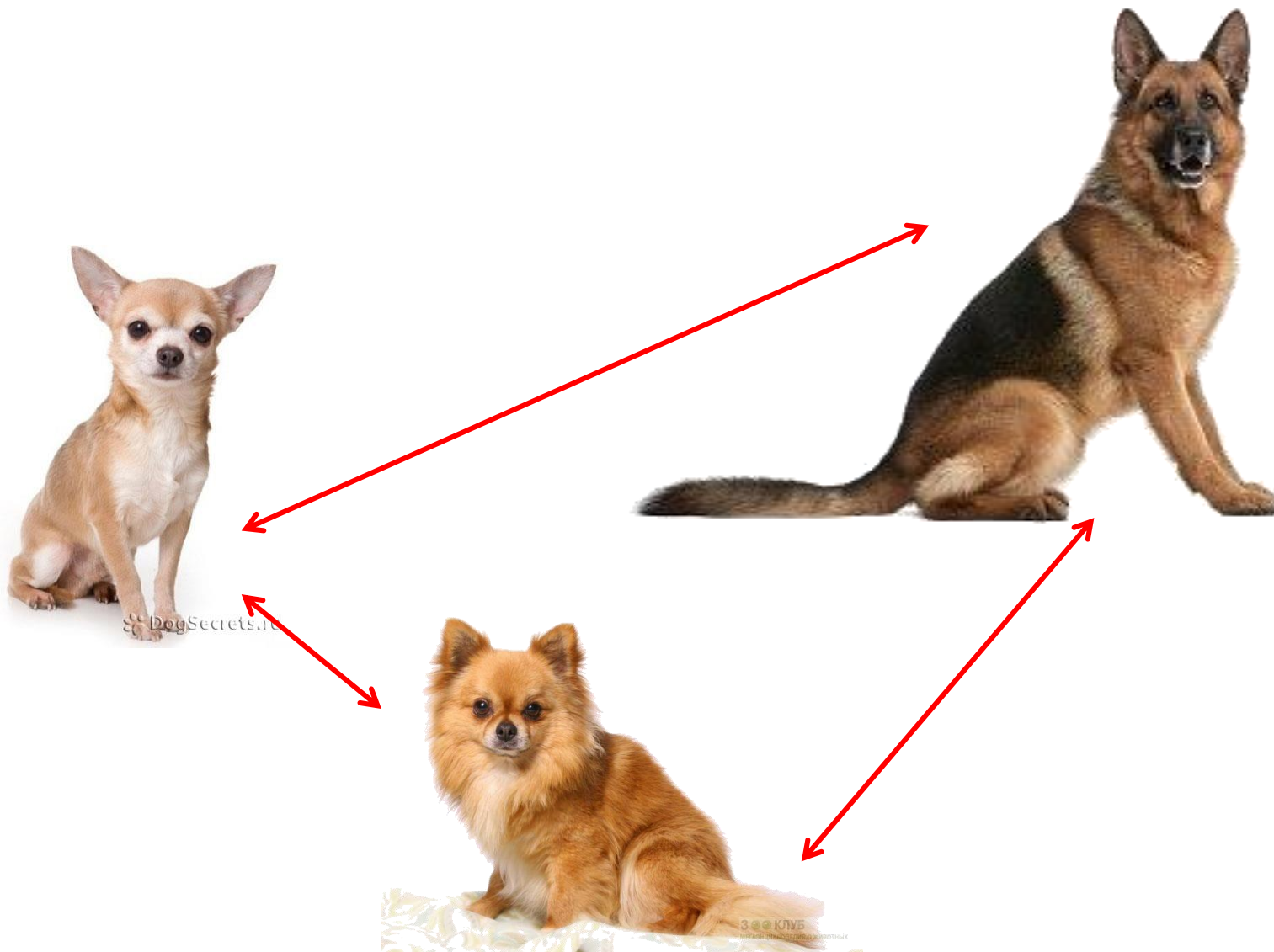
Матрица «дистанций» между объектами (dissimilarity matrix):



«Дистанции» между объектами показывают, насколько сходны между собой объекты (в парах) по всем переменным.

## Многомерный анализ

Матрица «дистанций» между объектами (dissimilarity matrix):



## Многомерный анализ

Есть много показателей «дистанции» между объектами (самый очевидный – **евклидовы расстояния**).

$$\sqrt{\sum_{j=1}^p (y_{1j} - y_{2j})^2}$$

Дистанции можно посчитать между объектами с любыми переменными, в т.ч. качественными и даже бинарными!

Это более демократичная основа для анализа, к ней перейдём в лекции 10.

### Подготовка данных для многомерного анализа

- ✓ первый этап - проверка данных на соответствие **нормальному** распределению и **линейность** связей – построение картинок (скаттерплотов и гистограмм);
- ✓ строим матрицу корреляций всех переменных между собой, ищем **сильно коррелирующие** и исключаем по одной из пары (уж точно если  $r > 0.9$ );
- ✓ **трансформация** данных: нормализует распределения и делает отношения между переменными **линейными** (важно для выделения компонент). Логарифмическая, квадратного корня и пр.
- ✓ важно избавиться от многомерных **аутлаеров**! После просмотра стандартных картинок, их можно найти с помощью дистанций Махаланобиса (квадрат расстояния от объекта до центроида); иногда от них помогает трансформация;

### Подготовка данных для многомерного анализа

- ✓ **стандартизация** данных: обязательна, если переменные измерены в принципиально разных шкалах, и различия в их значениях не имеют биологического смысла;
- ✓ для сравнения объектов можно предварительно построить **картинки** и оценить сходство и различие между объектами (лица Чернова, «звёздный» график);
- ✓ если для каких-то переменных есть пропущенные измерения, лучше выбирать не casewise, а pairwise deletion.

*Совет: попробовать проанализировать данные с разными вариантами трансформации/стандартизации.*

# Многомерный анализ



Лица Чернова

«звёздный» график –  
star plot

2D Graphs – Scatter Icon Plots, если объектов не очень много

# Сравнение групп объектов

Пусть мы имеем **МНОГОМЕРНЫЕ** данные. И они классифицируются на **ГРУППЫ** (какой-то группирующей переменной или переменными).

*Как сравнить группы между собой?*



*Примеры:*

- ✓ Мы имеем несколько морфологических промеров для зверьков; хотим сравнить особей разного возраста.
- ✓ Измерили разные физиологические показатели у разных видов растений; хотим сравнить теневыносливые и светолюбивые виды.

## MANOVA

Если бы зависимая переменная была одна, использовали бы ANOVA.

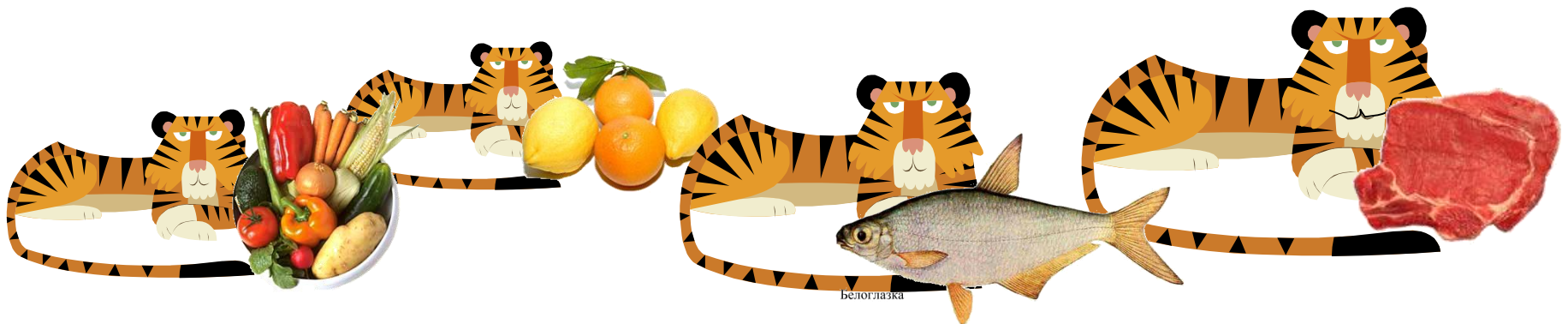
Почему бы не провести **отдельные** дисперсионные анализы для каждой из переменных?

1. Вероятность **ошибки 1-го рода** превысит **5%**;
2. Не будет учтена возможная **корреляция** между переменными;
3. Средние различия групп по каждой переменной могут быть малы, но по всем переменным **совместно** различия могут быть очевидными.

# MANOVA

Мы сравниваем **4 группы** тигров, у которых разный рацион; **зависимые переменные**: масса, упитанность, уровень кортикостероидов в крови.

$H_0$ : о влиянии группирующей переменной на **комбинацию зависимых** переменных = о равенстве центроидов в группах.



## Принципы MANOVA

1. Основа MANOVA – получение новой переменной - **линейную комбинации** зависимых переменных.
2. Группирующих переменных может быть **несколько** (одновременно и Muliway, и Multivariate дизайн)
3. Будем сравнивать **внутригрупповую** и **межгрупповую** дисперсии.
4. Разброс в многомерных данных характеризуется с помощью **матриц**, и MANOVA строит матрицы (SSCP) для отклонений **между группами** и **внутри групп**.

$$z_{ik} = c_1 y_{i1} + c_2 y_{i2} + \dots + c_j y_{ij} + \dots + c_p y_{ip}$$

## MANOVA

5. С помощью алгебры у этих двух матриц (SSCP межгрупповых и внутригрупповых) считается «отношение», и сразу получаются коэффициенты (eigenvectors) и собственные значения (eigenvalues, в программе - roots) для новых переменных;
6. линейная комбинация с максимальным eigenvalue при таком подходе получается такая, что для неё отношение межгрупповой и внутригрупповой дисперсий максимальное, т.е., для неё **различия между группами максимальны**.
7. Эту линейную комбинацию выбирают; она называется **дискриминантная функция** (discriminant function).

Эти манипуляции (5-7) в программе в стандартных MANOVA **не обозначены!** Там сразу просто идёт тестирование  $H_0$ .

# MANOVA

## Тестирование $H_0$ в MANOVA:

При помощи матриц SSCP между группами, внутри групп (аналог SS в ANOVA) и общей **тестируют гипотезу** об отсутствии различий между группами, для чего есть **несколько статистик**:

**Wilk's lambda** (отношение определителей **внутри**групповой SSCP и **общей** SSCP), чем она меньше, тем больше межгрупповые различия;

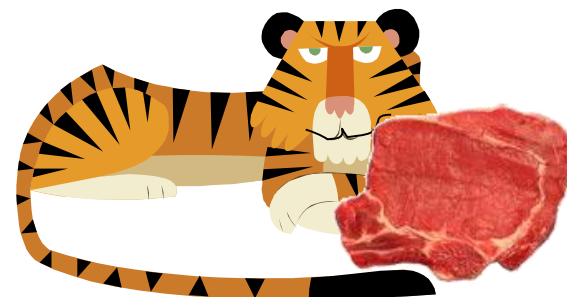
**Hotelling trace** (отношение определителей **меж**групповой SSCP и **внутри**групповой) — чем больше, тем больше различия групп;

**Pillai's trace** (сумма элементов главной диагонали — след — матрицы-отношения межгрупповой и общей SSCP), наиболее устойчив к отклонениям от многомерного нормального распределения и гомогенности дисперсии.

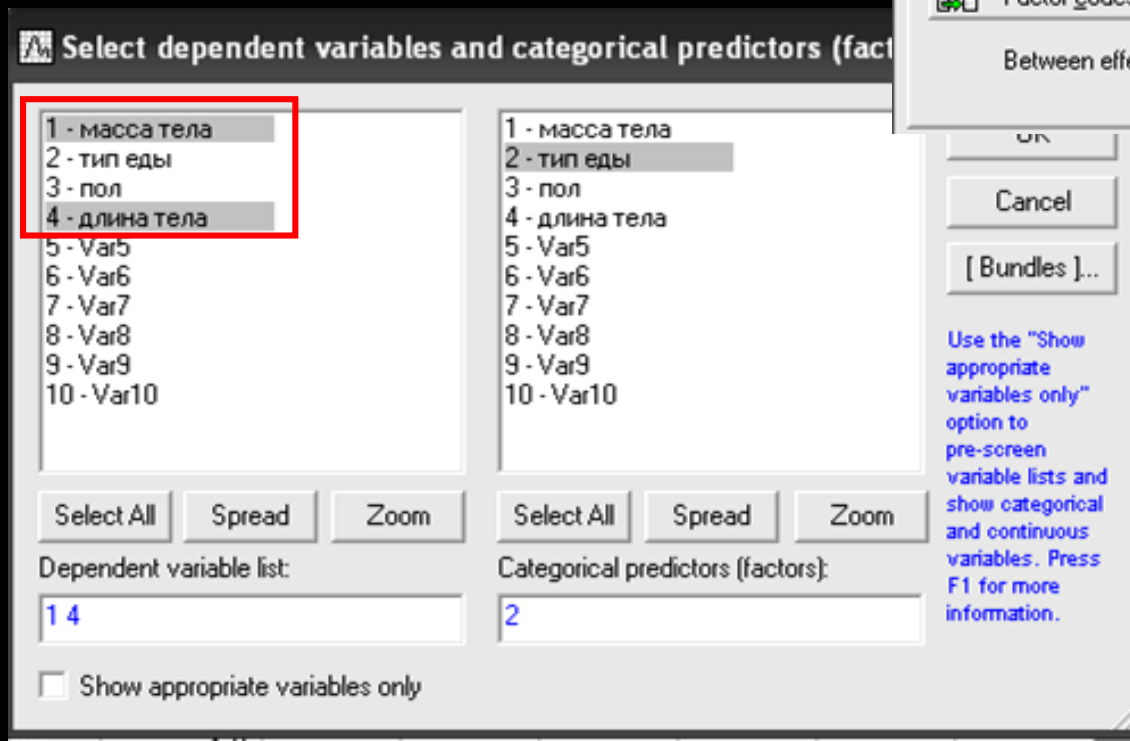
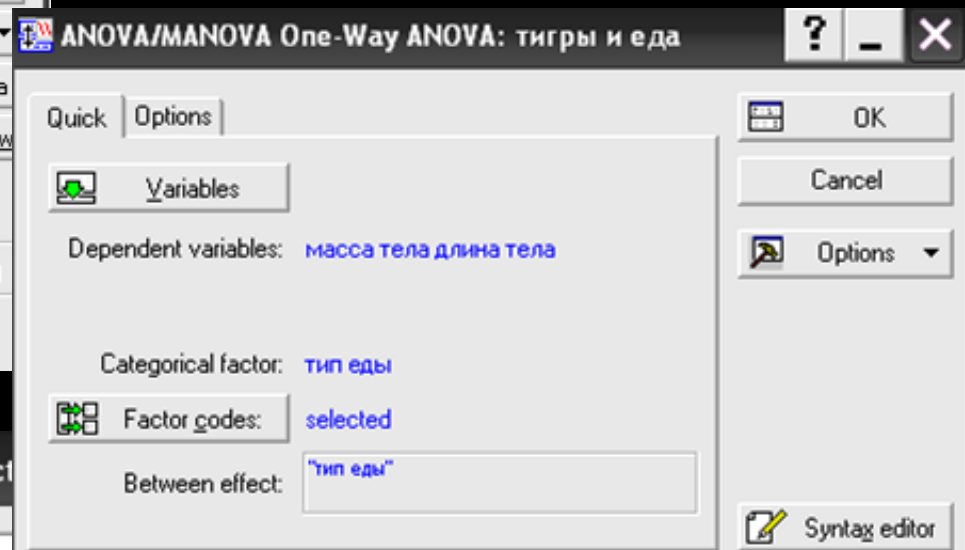
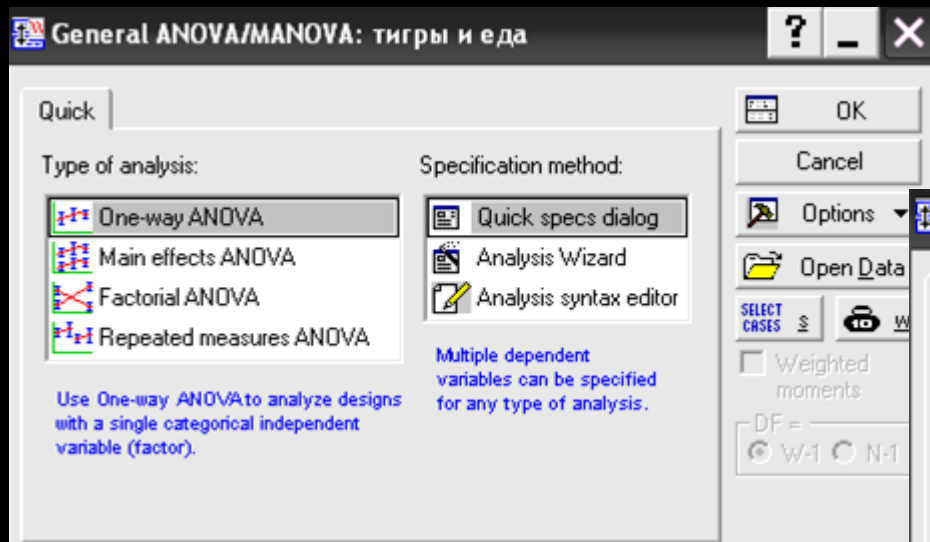
# MANOVA

- ✓ Все эти статистики преобразуют в величину, аппроксимирующуюся **F**-распределением (и их сравнивают с критическим F-значением).
- ✓ Если гипотеза **отвергнута**, проводят **POST-HOC ТЕСТЫ**
- ✓ Можно провести отдельные **univariate ANOVA**, чтобы понять, какие переменные имеют значения при разделении групп.

MANOVA может быть многофакторной, и можно проанализировать взаимодействие факторов



# MANOVA



# MANOVA

**ANOVA Results 2: тигры и еда**

Profiler | Resids | Matrix | Report  
Quick | Summary | Means | Comps

All effects/Graphs | All effects  
Univariate results | Cell statistics

Between effects  
Design terms | Whole model R  
Coefficients | Estimate

**Multivariate Tests of Significance (тигры и еда)**

Multivariate Tests of Significance (тигры и еда)  
Sigma-restricted parameterization  
Effective hypothesis decomposition

Effect	Test	Value	F	Effect df	Error df	p
Intercept	Wilks	0,002103	14000,64	2	59	0,00
тип еды	Wilks	0,049475	68,75	6	118	0,00

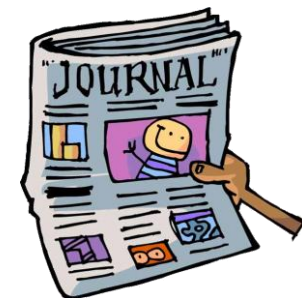
Multivariate Tests of Significance (тигры и еда)

**Univariate Results for Each DV (тигры и еда)**

Univariate Results for Each DV (тигры и еда)  
Sigma-restricted parameterization  
Effective hypothesis decomposition

Effect	Degr. of Freedom	масса тела SS	масса тела MS	масса тела F	масса тела p	длина тела SS	длина тела MS	длина тела F	длина тела p
Intercept	1	1012539	1012539	15798,81	0,00	3594342	3594342	16346,20	0,00
тип еды	3	42332	14111	220,17	0,00	67680	22560	102,60	0,00
Error	60	3845	64			13193	220		
Total	63	46177				80873			

тип еды; LS Means | тип еды; LS Means | Univariate Results for Each DV (тигры и еда)



*В методах:*

- ✓ не забыть указать, что распределение переменных соответствовало нормальному закону и между группами соблюдалась гомогенность дисперсии.
- ✓ указать, что пользовались многомерным дисперсионным анализом.

*For variables that conformed to a normal distribution (Shapiro–Wilk’s  $W$  test,  $p > 0.05$ ) and were homoscedastic (Levene’s test,  $p > 0.05$ ), we used multivariate analysis of variance (MANOVA) in general linear model (GLM).*

*В результатах:*

- ✓ Достаточно привести Wilk’s  $\lambda$  (или Pillai’s trace),  $F_{\text{effect df, error df}}$ ,  $p$ .

## Требования к выборкам для MANOVA

1. Многомерное **нормальное распределение**: довольно устойчива к отклонениям при одинаковых размерах групп, желательны одномерные нормальные распределения;
2. Очень чувствительна к **аутлаерам**.
3. Очень чувствительна к **гетерогенности дисперсий** в группах (достаточно проверить гомогенность для отдельных переменных).
4. Чем больше переменных в анализе, тем чувствительнее модель к нарушениям этих требований.
5. Не должно быть сильно **скоррелированных** переменных.
6. Очень желателен одинаковый размер групп

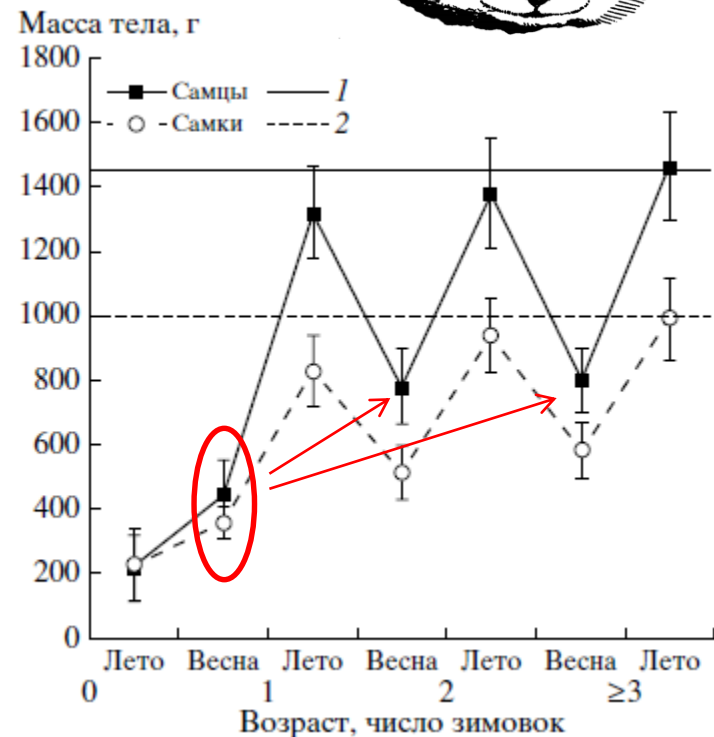
# Классификация объектов в группы

Задача очень похожа на сравнение групп объектов, но есть дополнительная цель: имея измеренные значения переменных, **классифицировать этот объект в ту или иную группу** (даже не зная её).



## *Пример из жизни сусликов:*

Как определять **возраст** у живых зверьков? Мы измерили у зверьков известного возраста: массу; ширину черепа; ширину резцов. Оказалось, что весной годовалые зверьки меньше, чем старшие. Потом отлавливая новых неизвестных особей, мы могли разделить их на годовалых и старших!



# ДИСКРИМИНАНТНЫЙ АНАЛИЗ (discriminant function analysis)

## Основная идея:

Мы измерили целый **НАБОР ПЕРЕМЕННЫХ**, и у нас **ИЗНАЧАЛЬНО** есть **ГРУППЫ** (одна группирующая переменная).

Мы хотим понять: 1) **чем отличаются** между собой эти группы (на основе данных переменных);

2) Насколько успешно на основе этих переменных мы можем **классифицировать** измерения в группы (скажем, когда мы потом измерим эти переменные у новой особи, мы сможем с известной вероятностью отнести её к той или иной группе).

## Дискриминантный анализ

Мы изучаем лемуров на Мадагаскаре.

У нас 3 вида лемуров, и мы хотим научиться определять вид зверька по черепам; мы в музее взвесили черепа, померили их длину и длину резцов.

**Вопрос:** на основе каких переменных отличаются виды и можем ли мы классифицировать особей по видам.



Нет возможности многофакторного анализа с оценкой взаимодействия факторов.

## Дискриминантный анализ

*Начинается как MANOVA, но имеет продолжение.*

**1. MANOVA:** на основе SSCP матриц (внутри и межгрупповой изменчивости) получаем **дискриминантные функции** и тестируем **гипотезу о различии** групп.

Дискриминантных функций **не больше**, чем число переменных или число групп -1 ( $\leq p-1$  или  $\leq k-1$ ); по первой из них группы лучше всего разделяются.

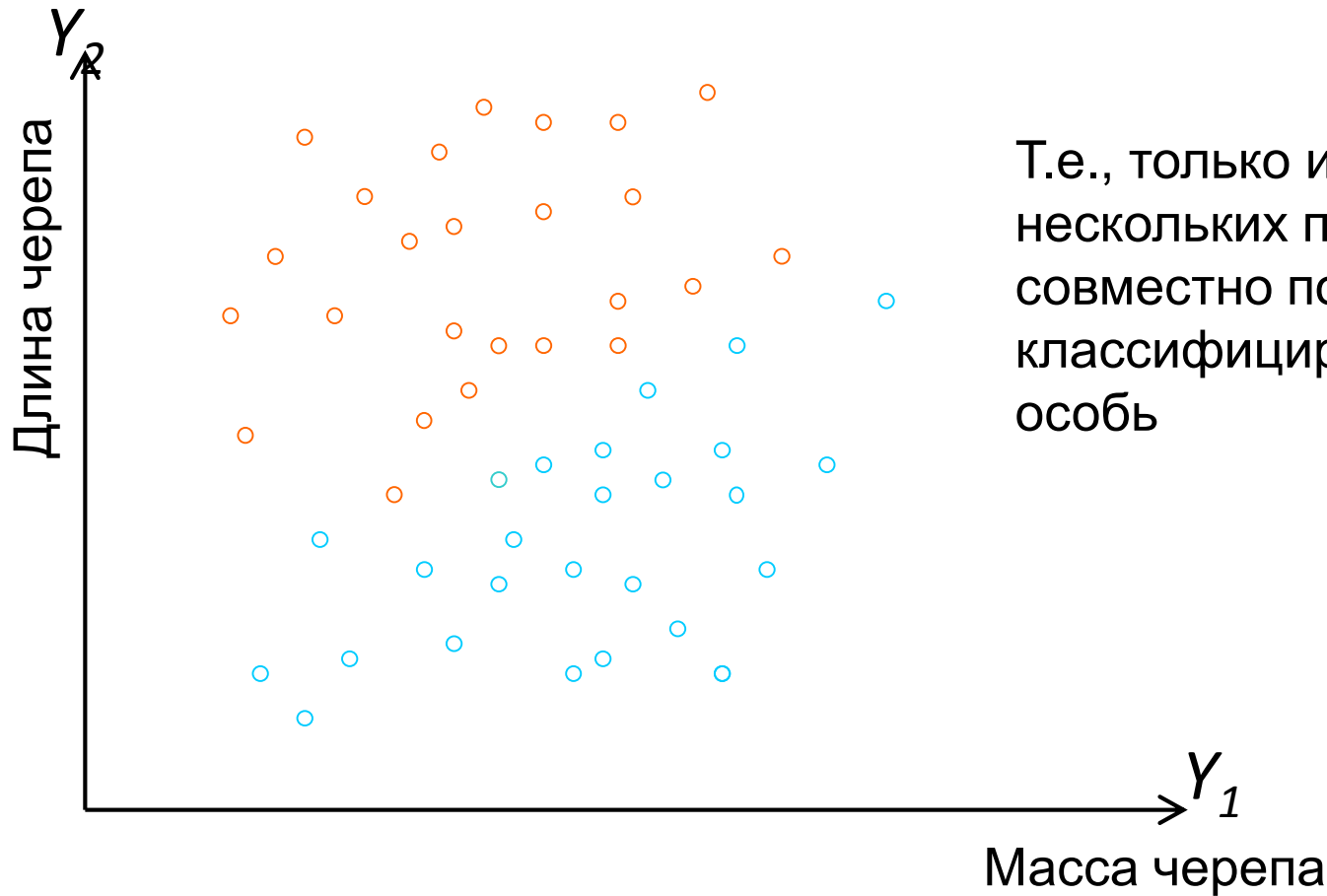
**2.** Если  $H_0$  отвергнута (различия есть), проверяем, **какие переменные** дают наибольший **вклад** в дискриминантные функции (смотрим на coefficients в функции и loadings=корреляции)

**3.** Можно провести пошаговый анализ и исключить не важные переменные

**4.** Получаем **классификационные функции** для каждой группы (в них мы будем подставлять наблюдаемые для объектов значения; объект запишем в ту группу, классификационная функция которой даст наибольшее значение)

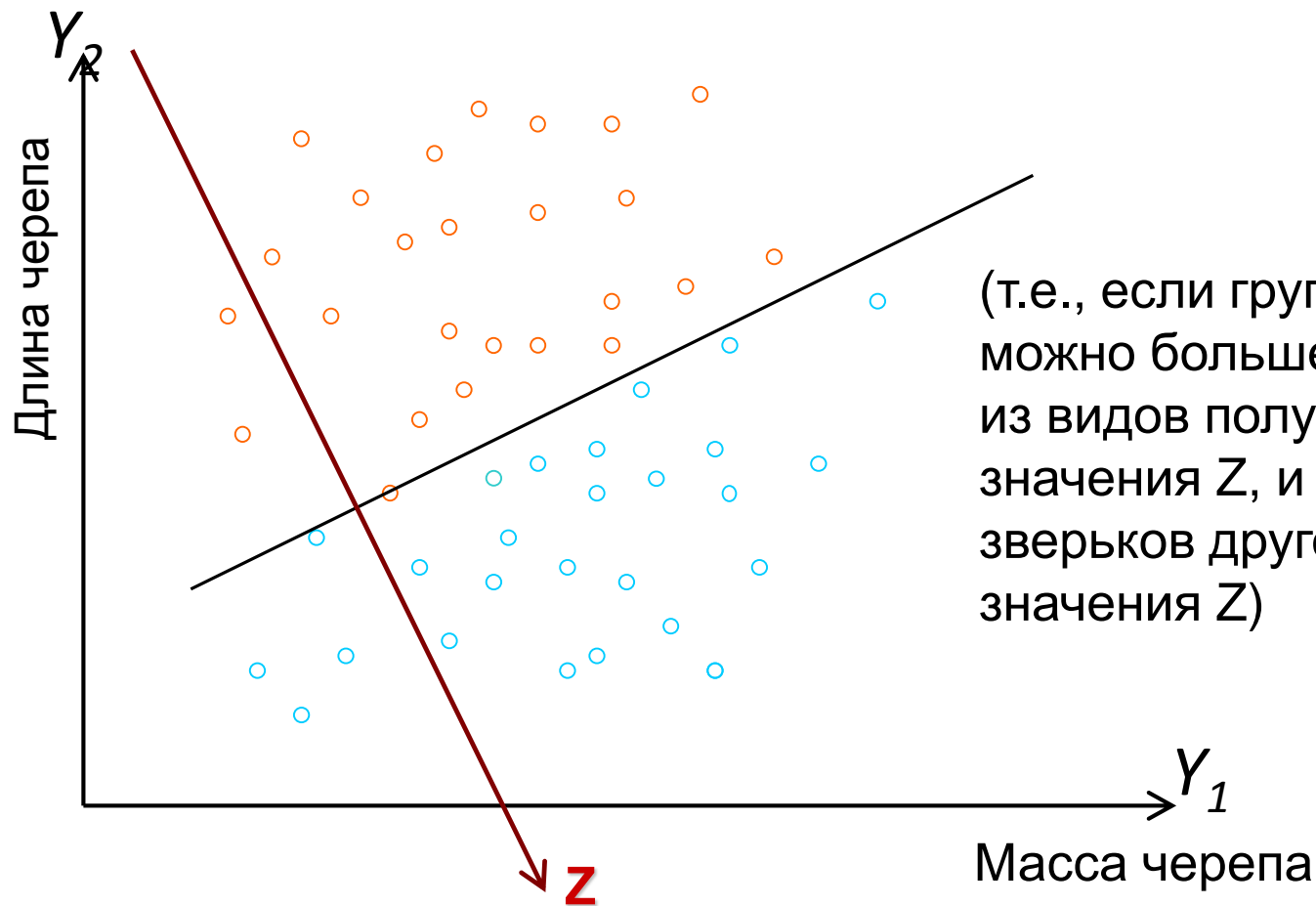
## Дискриминантный анализ

Предварительное рассмотрение скаттерплоттов: средние значения для каждой переменной у разных видов отличаются, но их распределения сильно перекрываются и для массы, и для головы, и для зубов.



## Дискриминантный анализ

Переменная  $Z$  (**дискриминантная функция**) получается такая, что если сравнить группы по этой функции, межгрупповая изменчивость у нее будет больше, чем у других.



(т.е., если группы 2, чтобы как можно больше зверьков одного из видов получили высокие значения  $Z$ , и как можно больше зверьков другого вида – низкие значения  $Z$ )

## Дискриминантный анализ

### Этап 1. Создание дискриминантной функции

Из исходных переменных рассчитываем **дискриминантные функции** – линейные комбинации исходных переменных, первая из которых наилучшим образом разделит группы (напр., виды). Вторая – «перпендикулярная» ей, на неё приходится максимум оставшейся межгрупповой изменчивости и т.п.

Если группы **две**: получается **одно** уравнение.

Когда групп и исходных переменных много, получают **несколько дискриминантных функций** (всего  $k-1$  или  $p-1$  функций,  $k$  – число групп,  $p$  – число переменных; выбирают меньшее из этих чисел), «перпендикулярных» друг другу.

$$z_{ik} = b_1 y_{i1} + b_2 y_{i2} + \dots + b_j y_{ij} + \dots + b_p y_{ip}$$

Тестируем  **$H_0$  о различии групп** в точности, как в MANOVA, используя ровно те же показатели

## Дискриминантный анализ

### Этап 2. Интерпретация дискриминантных функций

Каждую **дискриминантную функцию** характеризуют:

1. eigenvalue = **Root** (собственное значение), показывает, какую часть межгрупповой изменчивости объясняет функция. Можно проверить, сколько функций в модели действительно помогает различить группы, и исключить недостоверные.
2. eigenvector = **standardized b coefficients**,  $b_j$  – позволяют оценить вклад каждой из исходных переменных в данную дискриминантную функцию.

Структура факторов (**factor structure coefficients** = loadings) – позволяет понять, насколько какие переменные коррелируют с дискриминантными функциями.

If you want to assign substantive "meaningful" labels to the discriminant functions, then the structure coefficients should be used (interpreted); if you want to learn what is each variable's unique contribution to the discriminant function, use the discriminant function coefficients (weights).

## Дискриминантный анализ

### Этап 3. исключение «недостовверных» переменных - пошаговый анализ (необязательно)

**Смысл** – построить дискриминантную только из значимых переменных.

#### *Forward stepwise analysis:*

1. Переменные ранжируются по тому, насколько по ним хорошо различаются группы (в одномерном анализе).
2. Тестируется модель с самой лучшей переменной.
3. Тестируется модель (ANCOVA), где зависимая переменная – следующая по порядку, а та, что «лучше» неё добавлена как ковариата. Потом – следующая модель со следующей переменной, где ковариаты – «лучшие» переменные, и так пока различия между группами не перестанут быть значимыми.



На каждом шаге (для каждой переменной) считается статистика  $F$

## Дискриминантный анализ

### Этап 3. исключение «недостовверных» переменных - пошаговый анализ (необязательно)

**F to enter:** показывает, насколько хорошо группы отличаются по этой переменной в предварительном одномерном анализе (для Forward stepwise analysis)  
Можно задать минимальное значение, ниже которого переменная не будет включена в модель.

**F to remove:** то же самое; показывает, насколько «плохо» группы отличаются по этой переменной (для Backward stepwise analysis).

*Backward stepwise analysis:* начинают с модели, куда включены все переменные.



## Дискриминантный анализ

### Этап 4. Классификация

Из матриц (матрицы внутригрупповой изменчивости и матрицы средних значений всех переменных в каждой группе) получают **новые классификационные функции** (для каждой группы).

Подставляя в классификационные функции значения переменных для объекта, можно для него посчитать их значение (classification score) и отнести в ту или иную группу - **предсказать**, к какой группе относится особь, и оценить точность предсказания!

Можно провести на основе этих функций классификацию новых зверьков.

*Ещё раз:*

Дискриминантную функцию рассчитывают для объектов, изначально разделённых на группы.

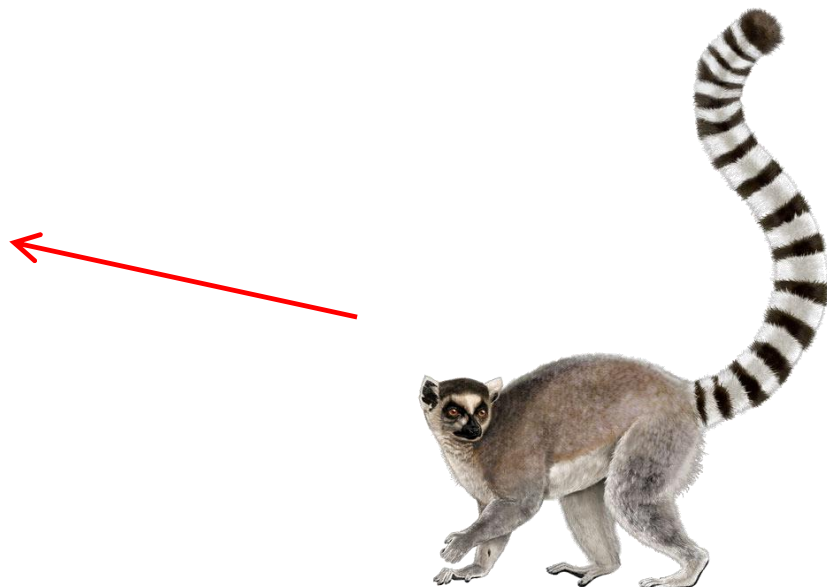


Если у нас есть набор признаков, и мы их на основе хотим **создать группы** (например, поделить вид на подвиды), это – **задача для другого анализа!**

## Дискриминантный анализ

Построив функции классификации, мы можем:

- ✓ поймать зверька неизвестного вида, измерить у него  $Y_1$ ,  $Y_2$ ,  $Y_3$ , рассчитать значения этих функций классификации, и с некоторой точностью **причислить** его к тому или другому виду;
- ✓ **проверить**, куда по этим функциям попадают зверьки, у которых группа известна (кто выпадает из своей группы и оказывается в чужой, и много ли таких).



# Discriminant function analysis

key HSD test; variable высота резца, см (данные к зачёту 25.03.sta)]

Statistics Data Mining Graphs Tools Data Workbook Window Scorecard Help

Resume... Ctrl+R

Basic Statistics/Tables

Multiple Regression

ANOVA

Nonparametrics

Distribution Fitting

Distributions & Simulation

Advanced Linear/Nonlinear Models

**Multivariate Exploratory Techniques**

Industrial Statistics & Six Sigma

Power Analysis

Automated Neural Networks

PLS, PCA, Multivariate/Batch SPC

Variance Estimation and Precision

Statistics of Block Data

STATISTICA Visual Basic

Batch (ByGroup) Analysis

Probability Calculator

Add to Report Add to MS Word Add to Worksp

м (данные к зачёту 25.03.sta)

key HSD test; variable высота резца, см (данные к зачёту 25.03.sta)

Proximate Probabilities for Post Hoc Tests

Source: Between MSE = ,00438, df = 96,000

возраст	{1}	{2}	{3}
	,53950	,33941	,54545

Cluster Analysis

Factor Analysis

Principal Components & Classification Analysis

Canonical Analysis

Reliability/Item Analysis

Classification Trees

Correspondence Analysis

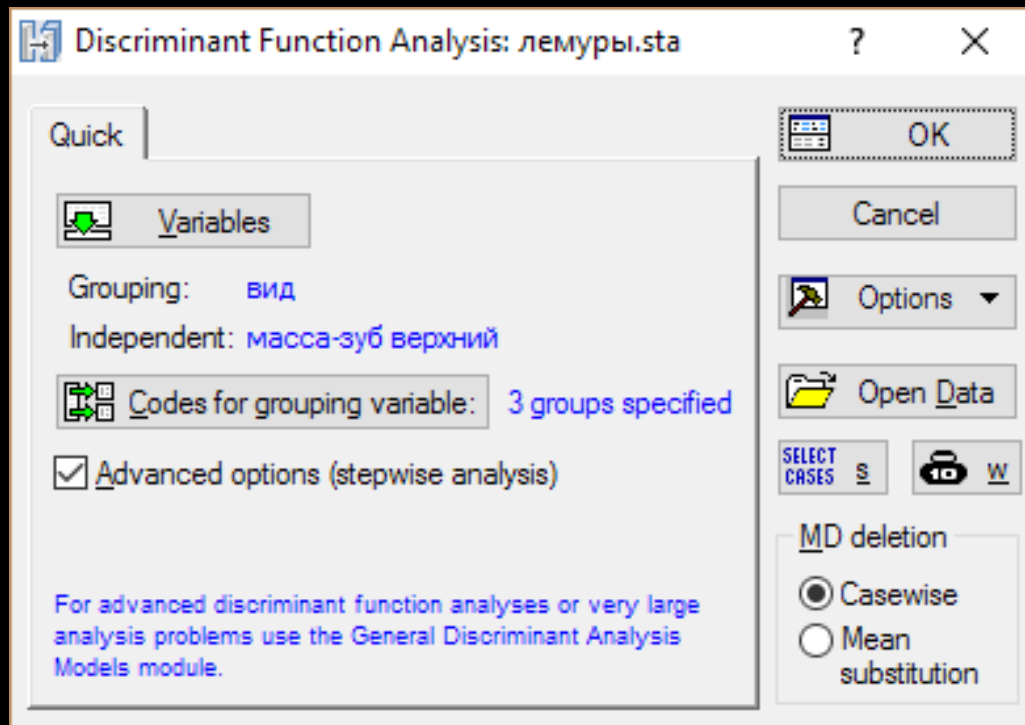
Multidimensional Scaling

**Discriminant Analysis**

General Discriminant Analysis Models

Data: лемуры.sta (9v by 58c)

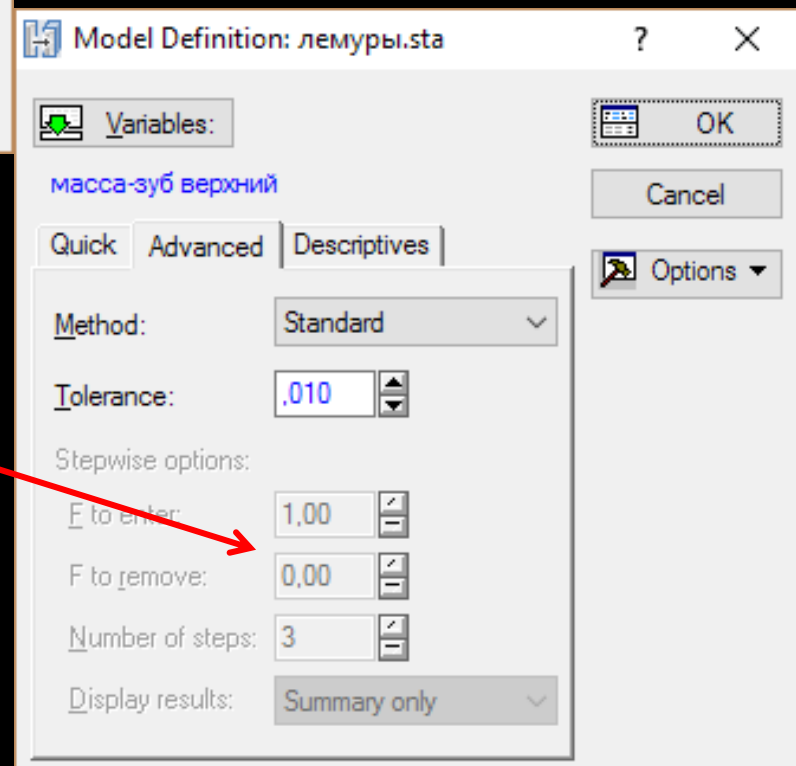
	1	2	3	4	5
	омер лемур	вид	масса	голова	зуб верхний
1	#430	кошачий	1848	59,4	4,5
2	#74	чёрный	4249	89	5,5
3	#291	сифака	3444	86	5,4
4	#461	кошачий	2442	62,25	4,8
5	#210	чёрный	3787	83,4	5,5
6	#1044	кошачий	1968	58,05	5
		сифака	3822	85,8	5,5
		чёрный	4746	89	5,7
		чёрный	4956	91	5,4
		сифака	2849	87,6	5,2
		сифака	4564	88,6	6,2
		чёрный	3857	85,2	6
		чёрный	4347	88,2	5,5
		сифака	3045	84	5,5
		чёрный	5509	90,8	6
		чёрный	3976	88,8	5,5
		сифака	3185	85,6	5,8
		чёрный	3836	88	5,4
		чёрный	3437	87,2	5,5
		кошачий	1956	58,5	4,8
		чёрный	4298	92	5,5
		сифака	3535	84	5,3

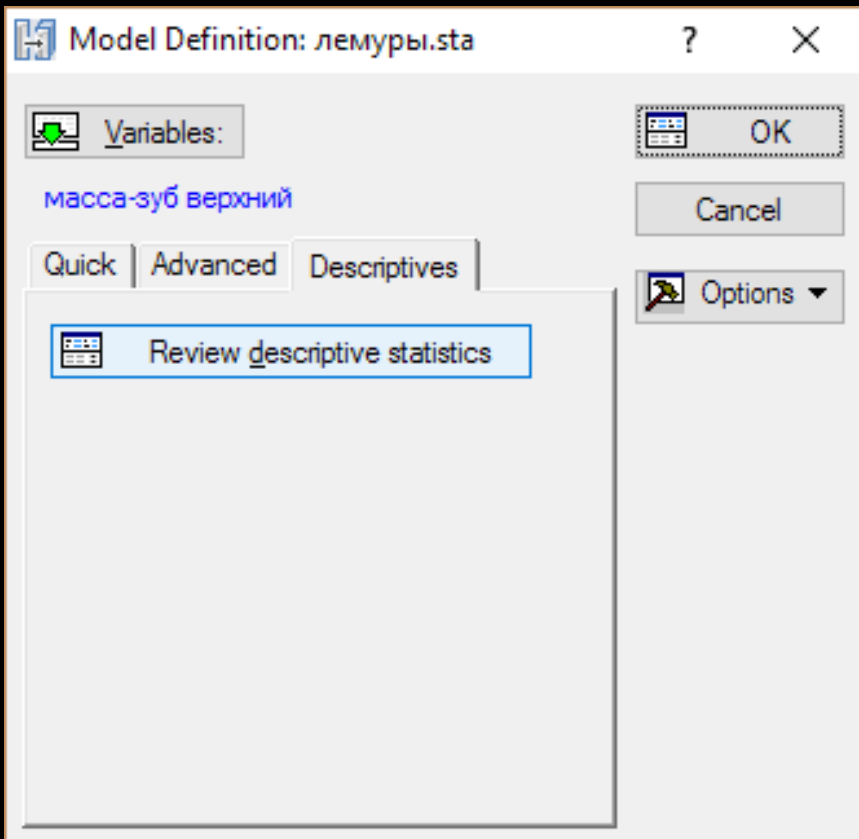


Выберем переменные для анализа.  
Выберем Advanced options чтобы исследовать данные предварительно

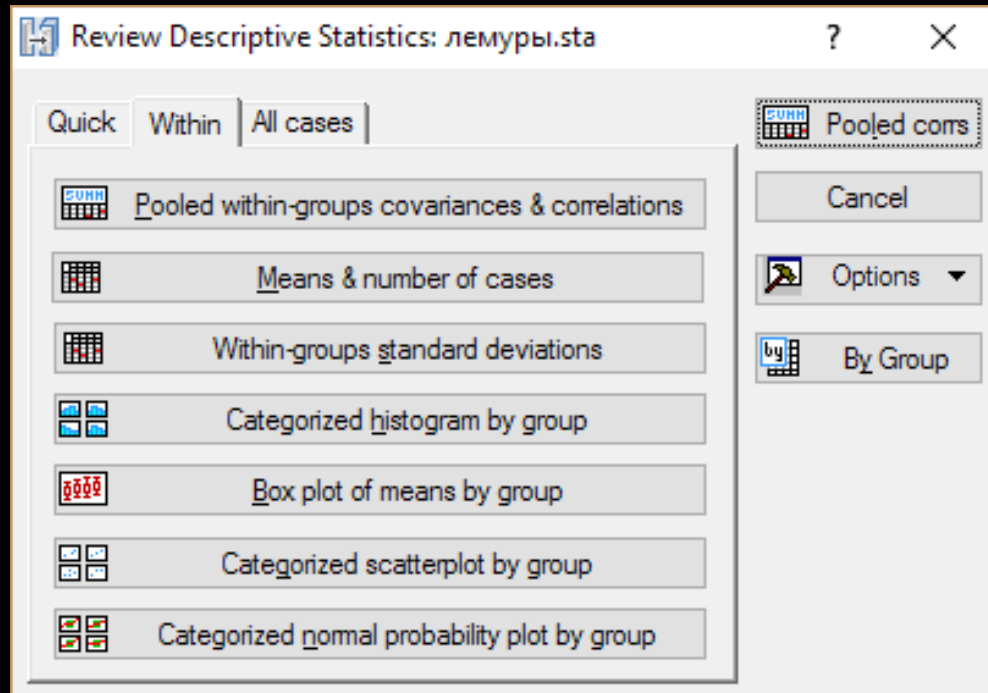
Критерии включения переменных в пошаговый анализ для построения дискриминантной функции. Лучше их задавать минимальными.

Толерантность —  $1 - R^2$ , где  $R^2$  оценивает корреляцию данной переменной с остальными, т.е., позволяет исключить избыточные переменные.



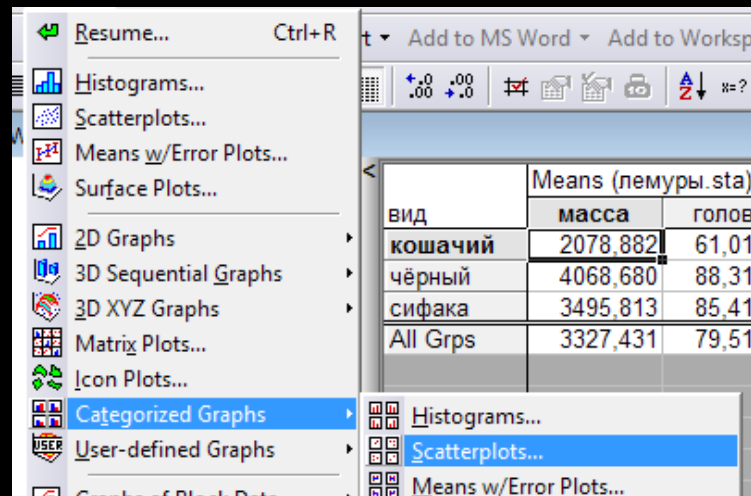


Можно исследовать переменные

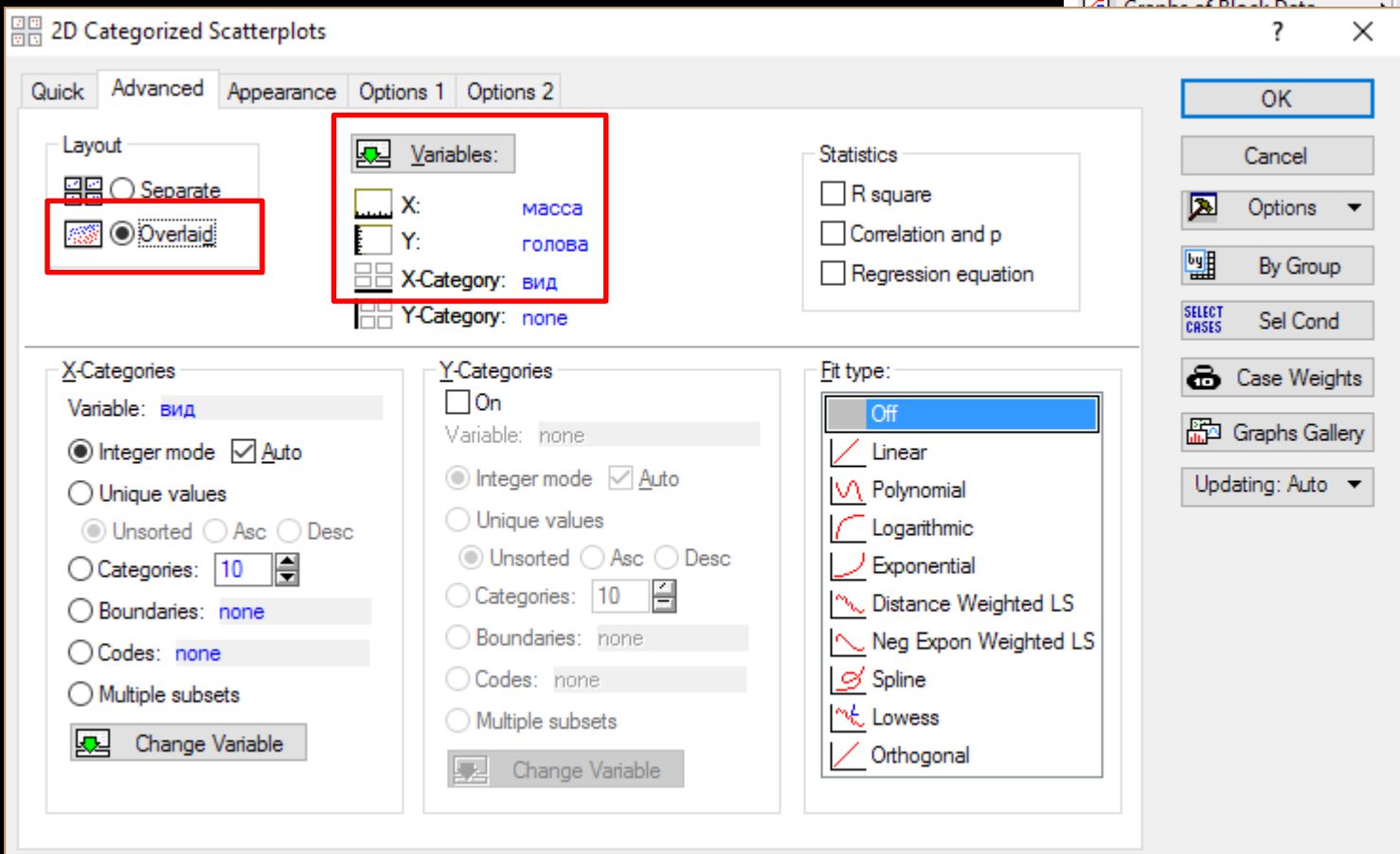
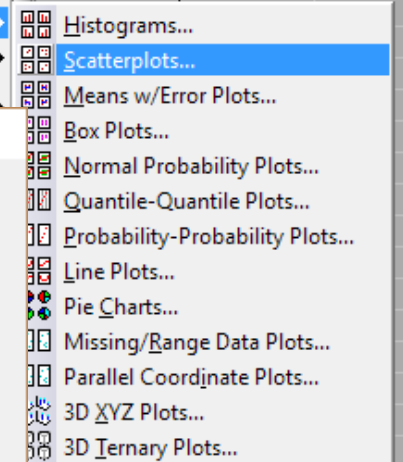


вид	Means (лемуры.sta)			
	масса	голова	зуб верхний	Valid N
кошачий	2078,882	61,01176	4,811765	17
чёрный	4068,680	88,31200	5,456000	25
сифака	3495,813	85,41250	5,337500	16
All Grps	3327,431	79,51035	5,234483	58

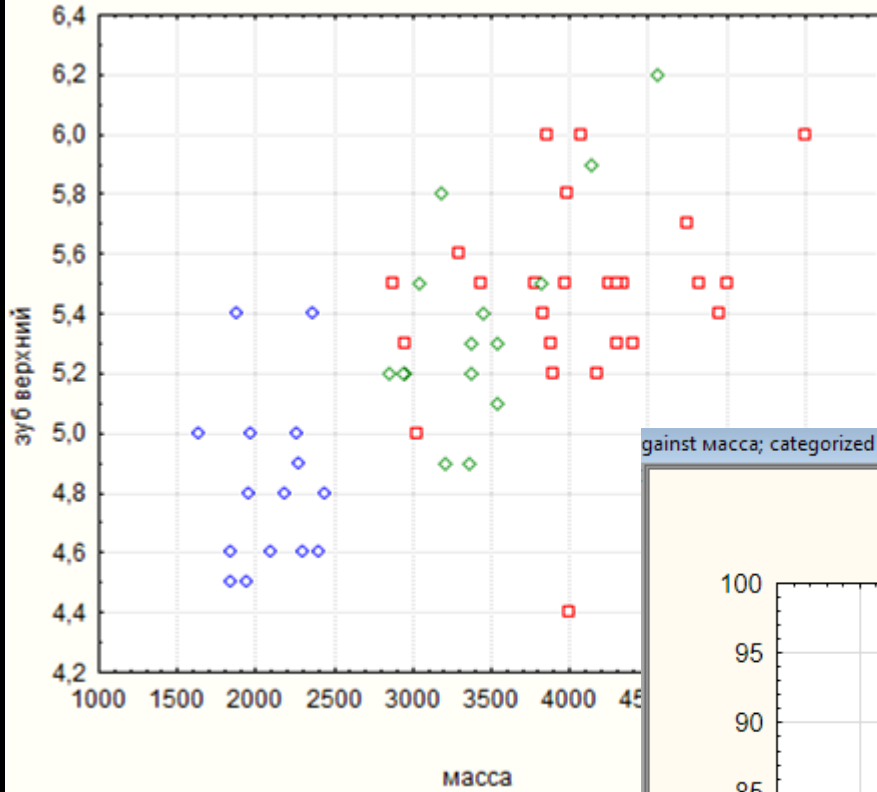
# Полезно построить картинки перекрывания групп по исходным переменным



Means (немурь.ста)		
вид	масса	голова
кошачий	2078.882	61.01
чёрный	4068.680	88.31
сифака	3495.813	85.41
All Grps	3327.431	79.51

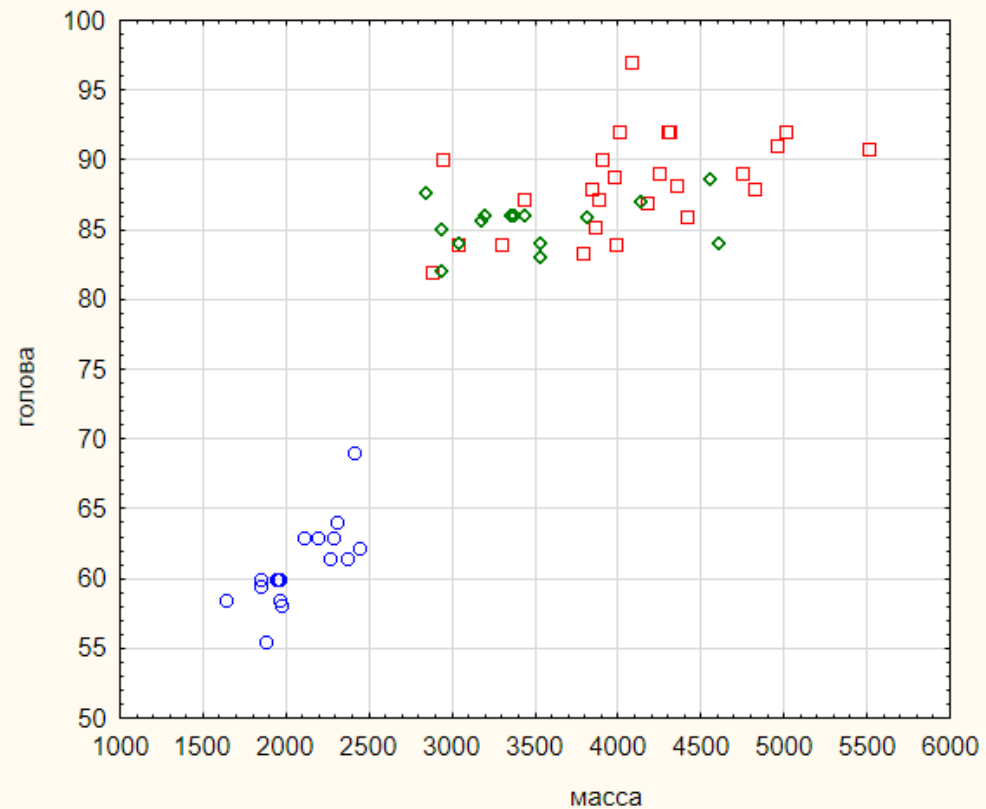


# Перекрывание групп



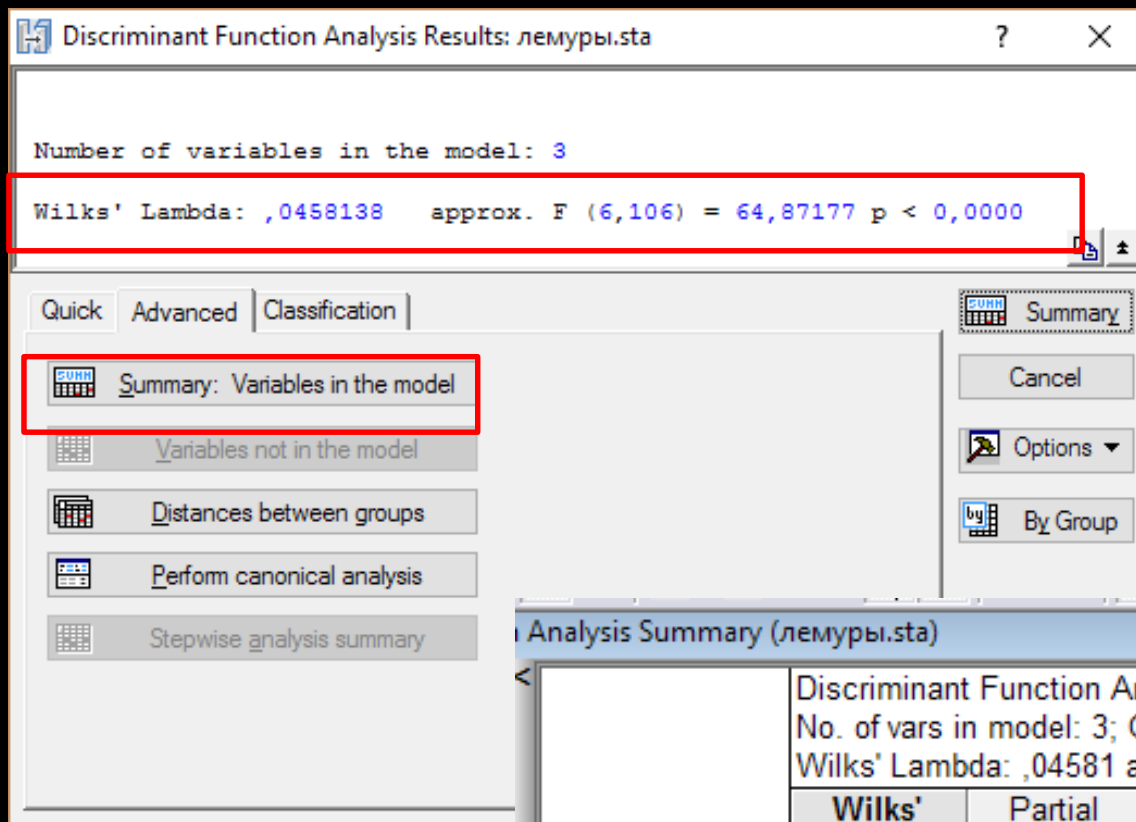
gainst масса; categorized by вид

Scatterplot of голова against масса; categorized by вид  
лемуры.sta 9v\*58c



○ вид: кошачий  
 □ вид: чёрный  
 ◇ вид: сифака





В целом, группы различаются

Analysis Summary (лемуры.sta)						
Discriminant Function Analysis Summary (лемуры.sta)						
No. of vars in model: 3; Grouping: вид (3 grps)						
Wilks' Lambda: ,04581 approx. F (6,106)=64,872 p<0,0000						
N=58	Wilks' Lambda	Partial Lambda	F-remove (2,53)	p-value	Toler.	1-Toler. (R-Sqr.)
масса	0,051783	0,884736	3,4524	0,038955	0,728760	0,271241
голова	0,257081	0,178208	122,2028	0,000000	0,775488	0,224512
зуб верхний	0,048547	0,943696	1,5811	0,215300	0,919741	0,080259

**Partial lambda** - статистика для оценки вклада каждой переменной в дискриминацию групп. Чем она меньше, тем больше вклад переменной. Переменная Голова лучше помогает различать виды, чем Масса.

**Wilk's lambda** — показывает, насколько хорошо будут различаться группы, если выкинуть переменную; чем меньше, тем меньше вклад.

# исследование дискриминантной функции

Дискриминантных  
функций 2, т.к. и групп,  
и переменных 3

Discriminant Function Analysis Results: лемуры.sta

Number of variables in the model: 3

Wilks' Lambda: ,0458138 approx. F (6,106) = 64,87177 p < 0,0000

Quick Advanced Classification

Summary: Variables in the model

Variables not in the model

Distances between groups

Perform canonical analysis

Canonical Analysis: лемуры.sta

Quick Advanced Canonical scores

Summary: Chi square tests of successive roots

Tests for canonical variables

Factor structure

of canonical variables

Summary

Cancel

Options

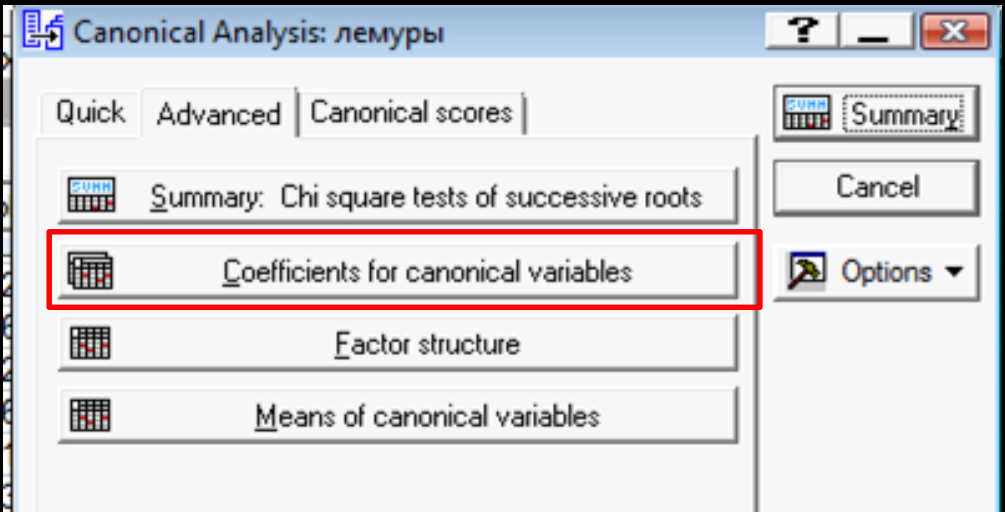
By Group

Square Tests with Successive Roots Removed (лемуры)

Chi-Square Tests with Successive Roots Removed (лемуры)						
Roots Removed	Eigen-value	Canonical R	Wilks' Lambda	Chi-Sqr.	df	p-level
0	18,62227	0,974186	0,045814	166,4911	6	0,000000
1	0,11238	0,317850	0,898972	5,7512	2	0,056382

Значимой оказалась только первая функция (root)

Посмотрим, какой вклад внесли переменные в различие групп нашими дискриминантными функциями.

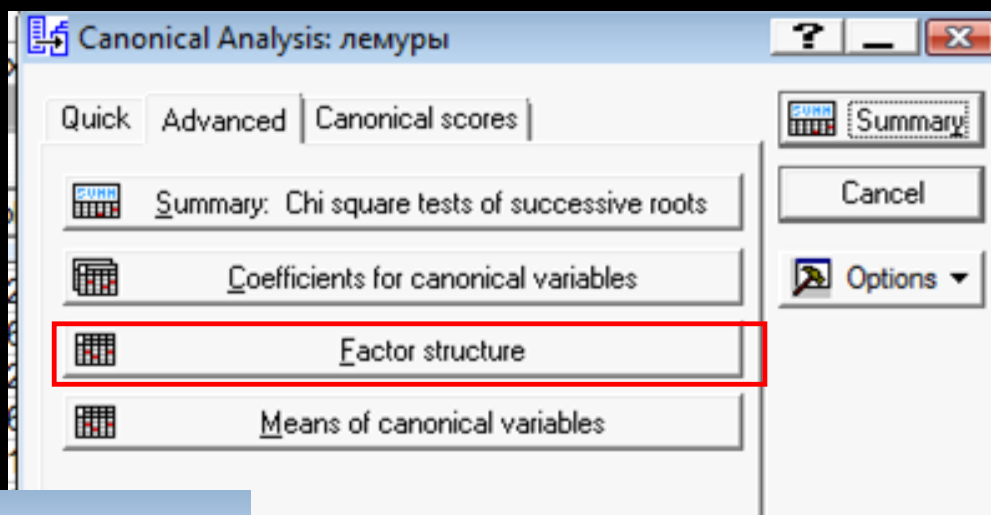


Standardized Coefficients (лемуры)			
Variable	Standardized Coefficients (лемуры) for Canonical Variables		
	Root 1	Root 2	
голова	-1,04774	0,42091	
масса	0,17011	-1,13742	
зуб верхний	-0,25299	0,06861	
Eigenval	18,62227	0,11238	
Cum.Prop	0,99400	1,00000	

*Standardized coefficients* – коэффициенты для сравнения значимости (**eigenvector**). «Голова» лучше всех позволяет различать группы

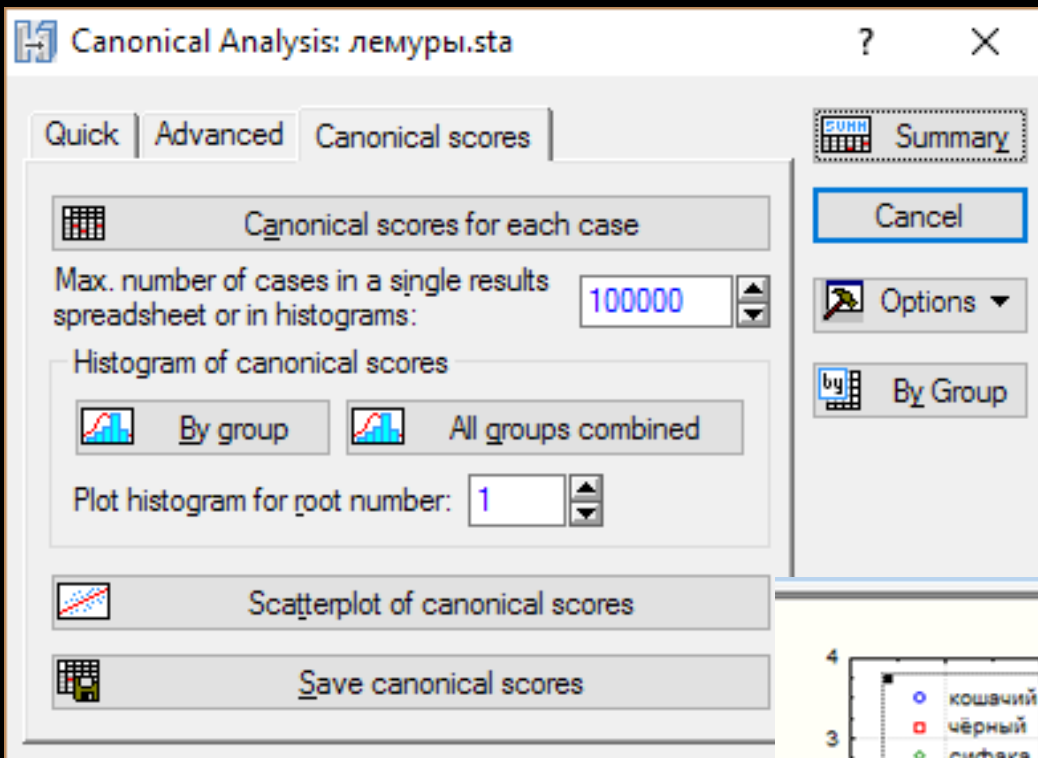
Первая функция объясняет 99,4% изменчивости

Структура факторов  
(дискриминантных  
функций) – **loadings**.



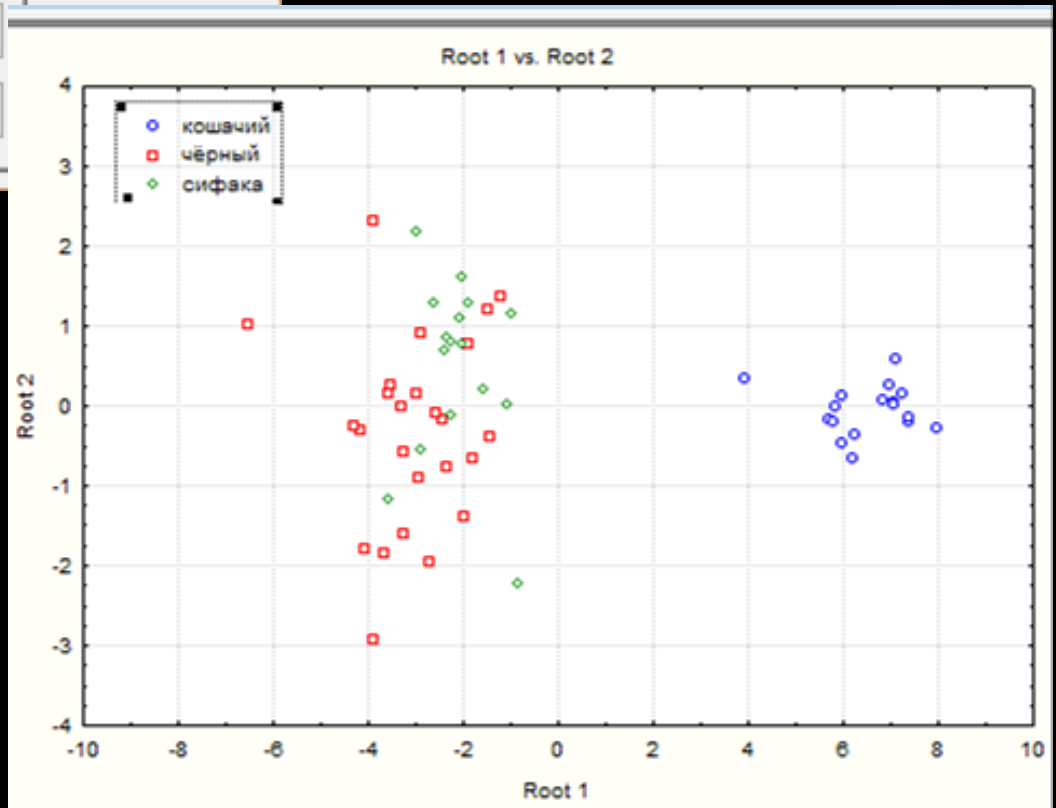
Factor Structure Matrix (лемуры)			
Correlations Variables - Canonical Roots (Pooled-within-groups correlations)			
Variable	Root 1	Root 2	
голова	-0,966831	-0,098483	
масса	-0,369679	-0,928690	
зуб верхний	-0,197236	-0,216579	

Наибольший вклад в первую функцию вносит Голова  
(она сильнее всего коррелирует с ней).

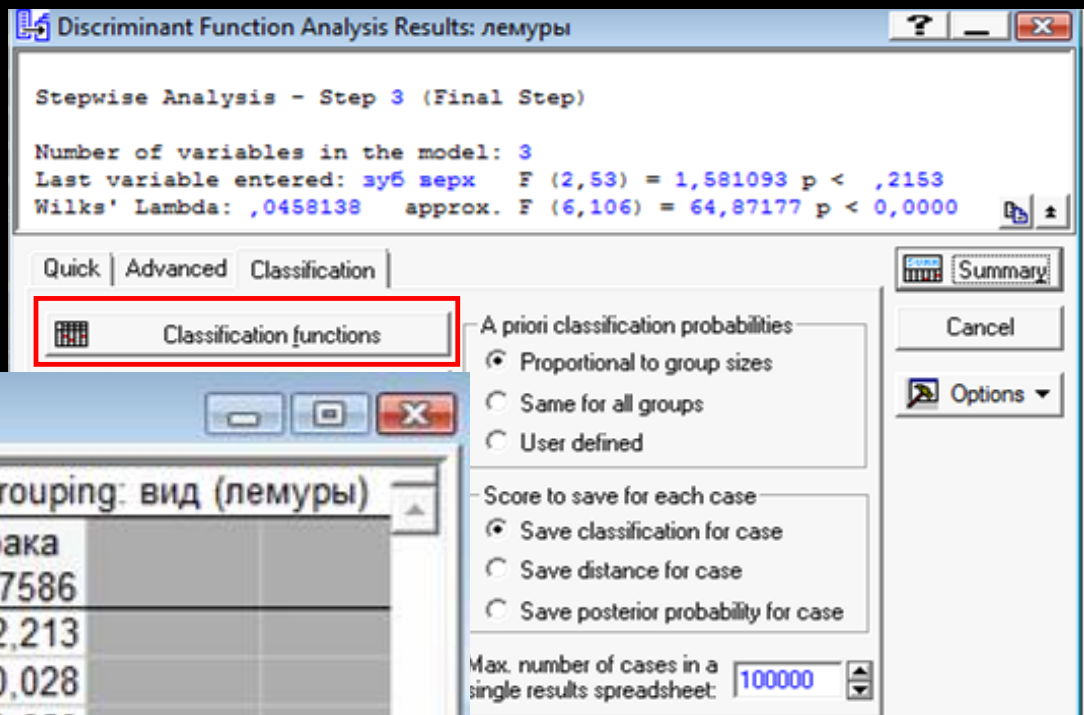


Мы можем посмотреть как располагаются виды в пространстве дискриминантных функций.

Кошачий лемур сильно отличается от других видов по значениям первой функции



# классификация



Classification Functions; grouping: вид (лемуры)

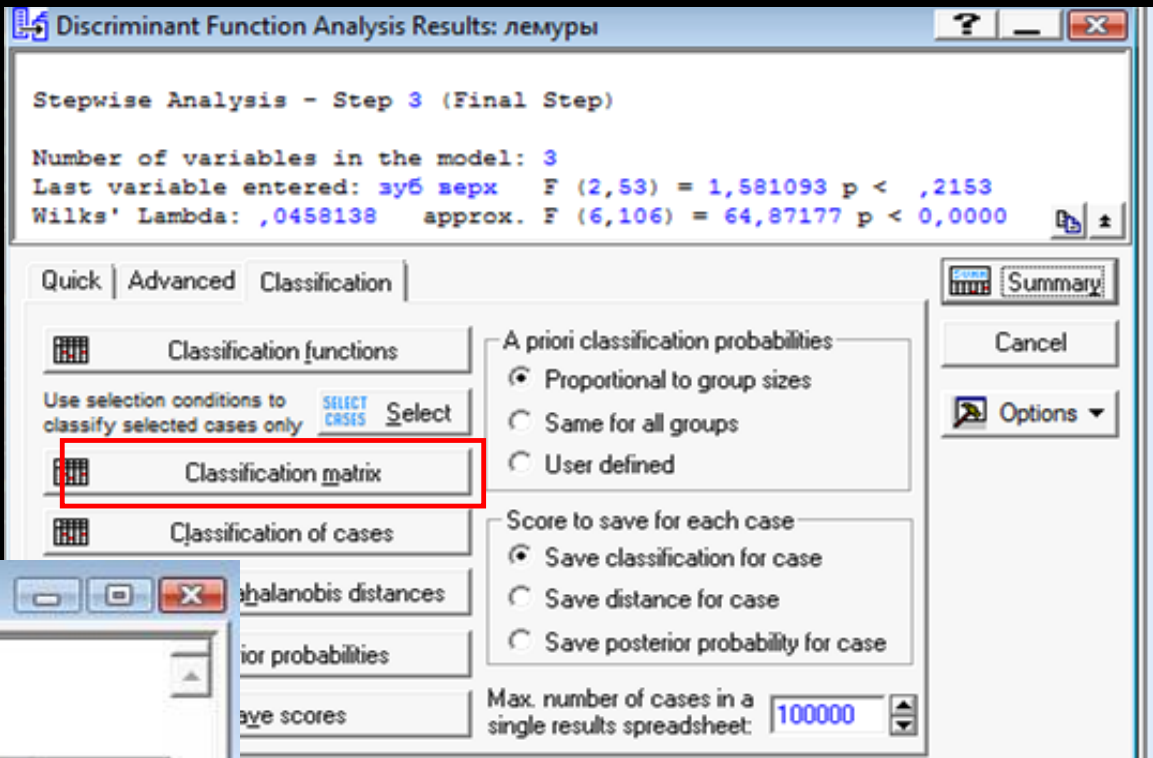
Variable	кошачий $p = ,29310$	чёрный $p = ,43103$	сифака $p = ,27586$
голова	9,068	12,433	12,213
масса	-0,024	-0,026	-0,028
зуб верхний	53,984	61,209	60,659
Constant	-382,876	-662,947	-636,011

Функции классификации : мы получаем для них коэффициенты, и можем классифицировать новых лемуров: взять новую особь, посчитать для неё функцию для каждой группы, и отнести её в ту группу, для которой значение будет наибольшим!

Значения  $p$  – вероятности случайного причисления лемура к той или иной группе, исходя из размеров группы.

Можно посмотреть, сколько лемуров правильно и неправильно причислено к той или иной группе на основе функций классификации.

## Классификационная матрица



Classification Matrix (лемуры)

Rows: Observed classifications  
Columns: Predicted classifications

Group	Percent Correct	кошачий p=,29310	чёрный p=,43103	сифака p=,27586
кошачий	100,0000	17	0	0
чёрный	80,0000	0	20	5
сифака	75,0000	0	4	12
Total	84,4828	17	24	17

Теперь можно взять других особей (они должны стоять в той же таблице) и посмотреть процент правильного причисления в группы

На основе дистанций Махаланобиса от каждого измерения до центра группы можно посмотреть, к какому виду тот или иной лемур причисляется. Неправильные причисления помечены звёздочками

Squared Mahalanobis Distances from Group Centroids (лемуры)

Case	Observed Classif.	кошачий p=.29310	чёрный p=.43103	сифака p=.27586
1	кошачий	1,1964	106,8041	88,3084
2	чёрный	95,9080	0,1279	2,4912
3	сифака	79,7296	1,4777	0,1531
4	кошачий	0,5235	85,6465	70,4008
* 5	чёрный	63,2996	2,8465	1,4498
6	кошачий	1,3327	109,4110	91,2075
* 7	сифака	77,1818	0,7576	0,5151
8	чёрный	98,0735	1,9369	5,9141
9	чёрный	107,3656	3,1768	8,2933
10	сифака	95,3058	6,1445	3,6353
* 11	сифака	106,7712	5,1732	9,0136
12	чёрный	83,3294	3,8958	4,0371
13	чёрный	90,2732	0,3758	2,6620
14	сифака	72,7316	4,4586	1,2447
15	чёрный	117,1906	8,4964	15,7356
16	чёрный	95,9827	0,1566	1,6663
17	сифака	87,6743	5,0135	3,2262
18	чёрный	89,9184	0,2069	0,8619
19	чёрный	89,4896	1,6823	0,9760

Discriminant Function Analysis Results: лемуры

Stepwise Analysis - Step 3 (Final Step)

Number of variables in the model: 3  
 Last variable entered: зуб верх F (2,53) = 1,581093 p < ,2153  
 Wilks' Lambda: ,0458138 approx. F (6,106) = 64,87177 p < 0,0000

Quick | Advanced | Classification | Summary

Classification functions  
 Use selection conditions to classify selected cases only SELECT CRISIS Select

Classification matrix

Classification of cases  
Squared Mahalanobis distances

Posterior probabilities

Save scores

A priori classification probabilities  
☒ Proportional to group sizes  
☐ Same for all groups  
☐ User defined

Score to save for each case  
☒ Save classification for case  
☐ Save distance for case  
☐ Save posterior probability for case

Max. number of cases in a single results spreadsheet: 100000

Cancel Options

## Требования к выборкам для дискриминантного анализа (как для MANOVA)

1. Многомерное нормальное распределение: довольно устойчив к отклонениям при **одинаковых размерах групп**, желательны одномерные нормальные распределения;
2. Очень чувствителен к **аутлаерам**
3. Очень чувствителен к **гетерогенности** дисперсий (**необходимо** проверить гомогенность для отдельных переменных)
4. Чем больше переменных в анализе, тем чувствительнее модель к нарушениям этих требований.
5. Не должно быть чрезмерно коррелирующих друг с другом переменных. Один из симптомов сильно скореллированных переменных (мультиколлинеарности) – несоответствие паттерна у коэффициентов и loadings (например, коэффициент у данной переменной большой, а loadings с этой функцией маленький).



## В публикацию



*Методы:* написать, что использовали дискриминантный анализ, что переменные соответствовали нормальному распределению и условию гомогенности; если пошаговый, указать, что Forward stepwise, P to enter.

*Результаты:* приводим общую Wilk's lambda, F, p; Eigenvalues и или Loadings (корреляции), или коэффициенты (из Standardized coefficients) – показателями вклада отдельных переменных. Иногда приводят матрицу классификации и процентом верного причисления.