Среда 12.12. 16:00
Конференц-зал ИЦИГ

Публичная лекция
В.М.Ефимова

Зачем БИОлогу БИОстатистика

Цель биостатистики

Адекватная математическая обработка
статистических данных
для решения биологических задач

Биологические задачи (примеры)

Исследование и описание изменчивости данной совокупности объектов

(например, инвентаризация населения птиц Западно-Сибирской равнины)

Исследование различий между совокупностями объектов по данной системе описаний

(например, по массе тела между контрольной и экспериментальной группой)

Исследование взаимосвязи между различными системами описаний объектов

(например, между генотипическими и фенотипическими признаками)

Выявление закономерностей наследования фенотипических свойств организмов

(например, выведение урожайных сортов ячменя, обеспечивающих высокое качество пива и устойчивых к колебаниям климата)

Данные

Совокупность описаний объектов

Способы получения данных

Эксперименты, наблюдения,
компьютерные симуляции

Основные типы данных

1. Таблица “объект–признак”
2. Матрица сходства/различия между объектами

Матрица сходства/различия между объектами может вычисляться по таблице “объект–признак” или просто задаваться пользователем.

Основные одномерные и многомерные методы анализа биологических данных

Классические линейные методы

+ кластерный анализ

Докомпьютерная эпоха

Нелинейные методы

(кроме кластерного анализа)

Эпоха “больших” компьютеров

Современные линейные методы

Эпоха супер- и персональных компьютеров

Классические линейные методы

Вычисление квантилей, средних, среднеквадратичных отклонений, дисперсий, коэффициентов корреляции,

ранжирование, центрирование и нормирование признаков, построение графиков,

линейная регрессия, множественная линейная регрессия,

дисперсионный анализ, дискриминантный анализ,

метод главных компонент, факторный анализ

Нелинейные методы

Вычисление коэффициентов сходства/различия (расстояний) между объектами, кластерный анализ,

карты Кохонена, многомерное шкалирование,

нелинейная регрессия, нейронные сети

Современные линейные методы

PLS-регрессия, 2B–PLS-регрессия,
дискриминантный PLS-анализ

Одномерная линейная регрессия

Ф.Гальтон
1822–1911

Исторически многомерный анализ биологических данных начался с работ Френсиса Гальтона (1822–1911), который попытался рассмотреть зависимость между средним ростом родителей и средним ростом их потомков. Предположив линейный характер зависимости и построив ее график по методу наименьших квадратов (что в те времена было совсем нетривиальным), он обнаружил, что потомки в среднем ближе к популяционной средней, чем родители. Гальтон назвал это явление "регрессией" и с тех пор так называется любая функциональная зависимость одной переменной от одной или нескольких других, подобранная статистическими методами.

Ф.Гальтон – двоюродный брат Ч.Дарвина. Открыл антициклоны, основал дактилоскопию, евгенику, психометрику, генетику количественных признаков и биометрию (1889).

Одномерная линейная регрессия

Ф.Гальтон
1822–1911

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

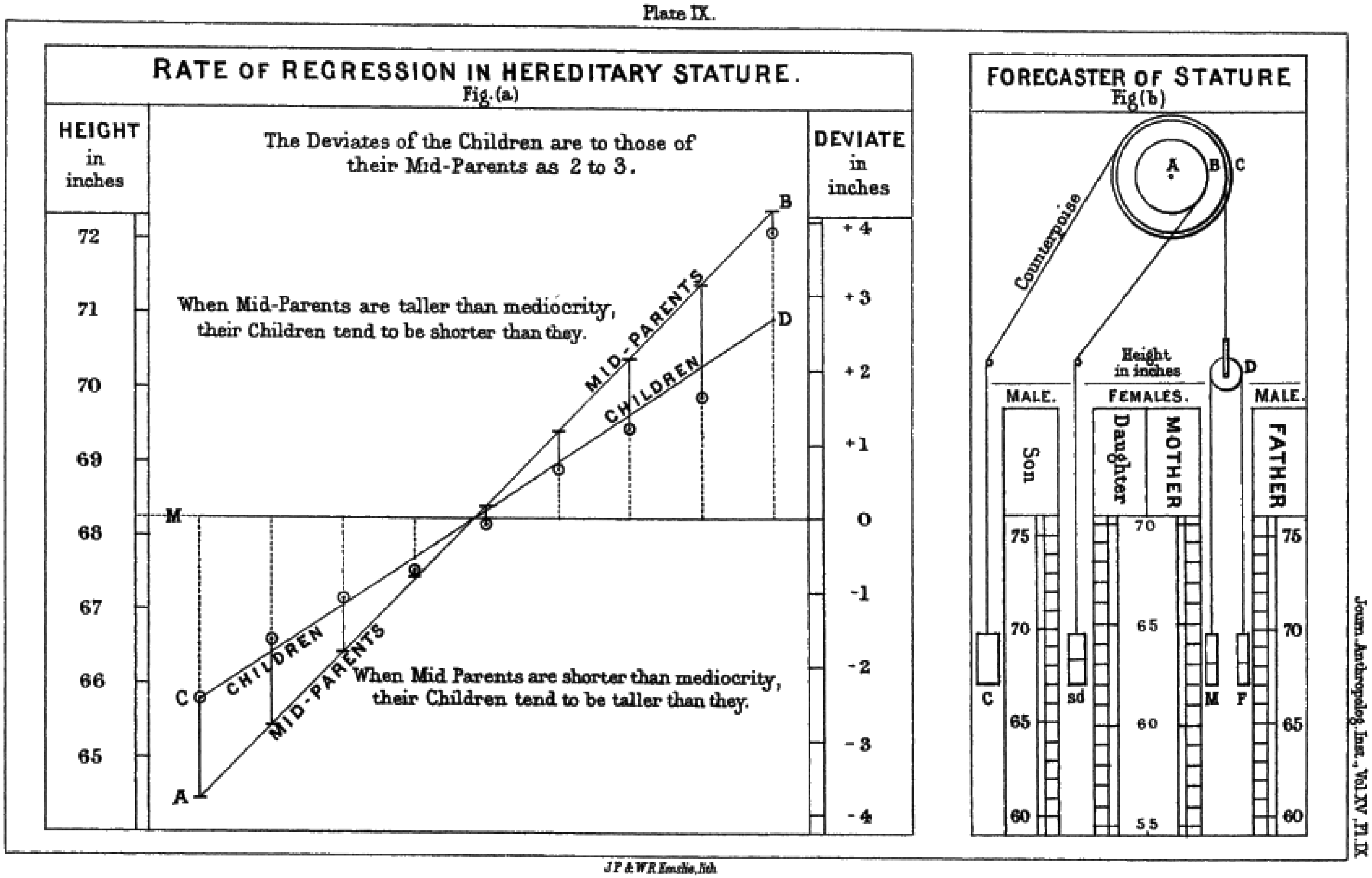
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Одномерная линейная регрессия

Ф.Гальтон
1822–1911



Одномерная линейная регрессия

Ф.Гальтон

Уравнение линейной регрессии. Метод наименьших квадратов

$$y = ax + b$$

$$y = r\frac{s_y}{s_x}x + \bar{y} - r\frac{s_y}{s_x}$$

$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}$$

$$y' = rx'$$

Множественная линейная регрессия

Карл Пирсон
1857—1936

Карл Пирсон (1857–1936) теоретически обосновал и разработал хорошо всем известный коэффициент линейной корреляции (коэффициент Браве–Пирсона) и много других коэффициентов, а также ввел понятие "множественной регрессии", т.е. функциональной зависимости одной переменной от нескольких других. Для определения параметров используется метод наименьших квадратов. Важнейшим частным случаем является множественная линейная регрессия. Кроме того, он первым предложил метод построения главных компонент (Pearson, 1901). Однако в то время на эту работу не обратили никакого внимания, да и сам Пирсон больше к ней не возвращался. Он же вместе с Уэлдоном и Гальтоном (консультант-редактор) основал журнал "Биометрика" для статистического изучения биологических проблем (1901).

Множественная линейная регрессия

К. Пирсон

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_m x_m + b$$

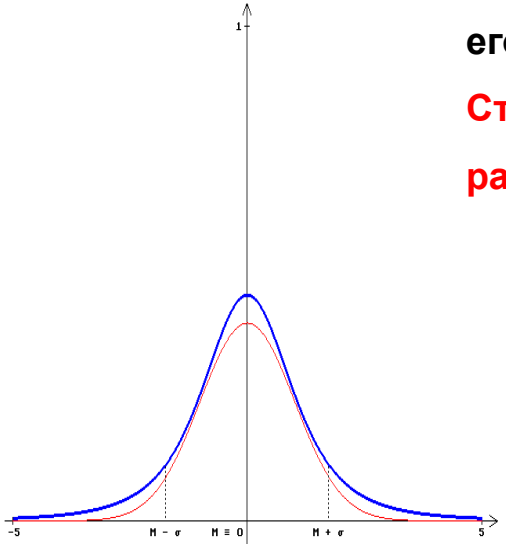
$$y = \sum a_i x_i + b$$

y – зависимая переменная

$x_1 \ldots x_m$ – независимые переменные

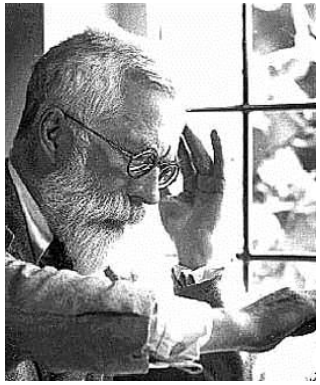
Распределение Стьюдента (Госсета)

Госсет, Уильям Сили
1876 - 1937



По окончании Оксфордского университета в 1899 году поступил на работу на пивоваренный завод Гиннесса. Чтобы предотвратить раскрытие конфиденциальной информации, Гиннесс запретил своим работникам публикацию любых материалов, независимо от содержавшейся в них информации. Это означало, что Госсет не мог опубликовать свои работы под своим именем. Поэтому он избрал себе псевдоним Стьюдент, и его самое важное открытие получило название “распределение Стьюдента”, иначе бы оно могло называться теперь распределением Госсета. Автор t-критерия Стьюдента.

Дискриминантный анализ



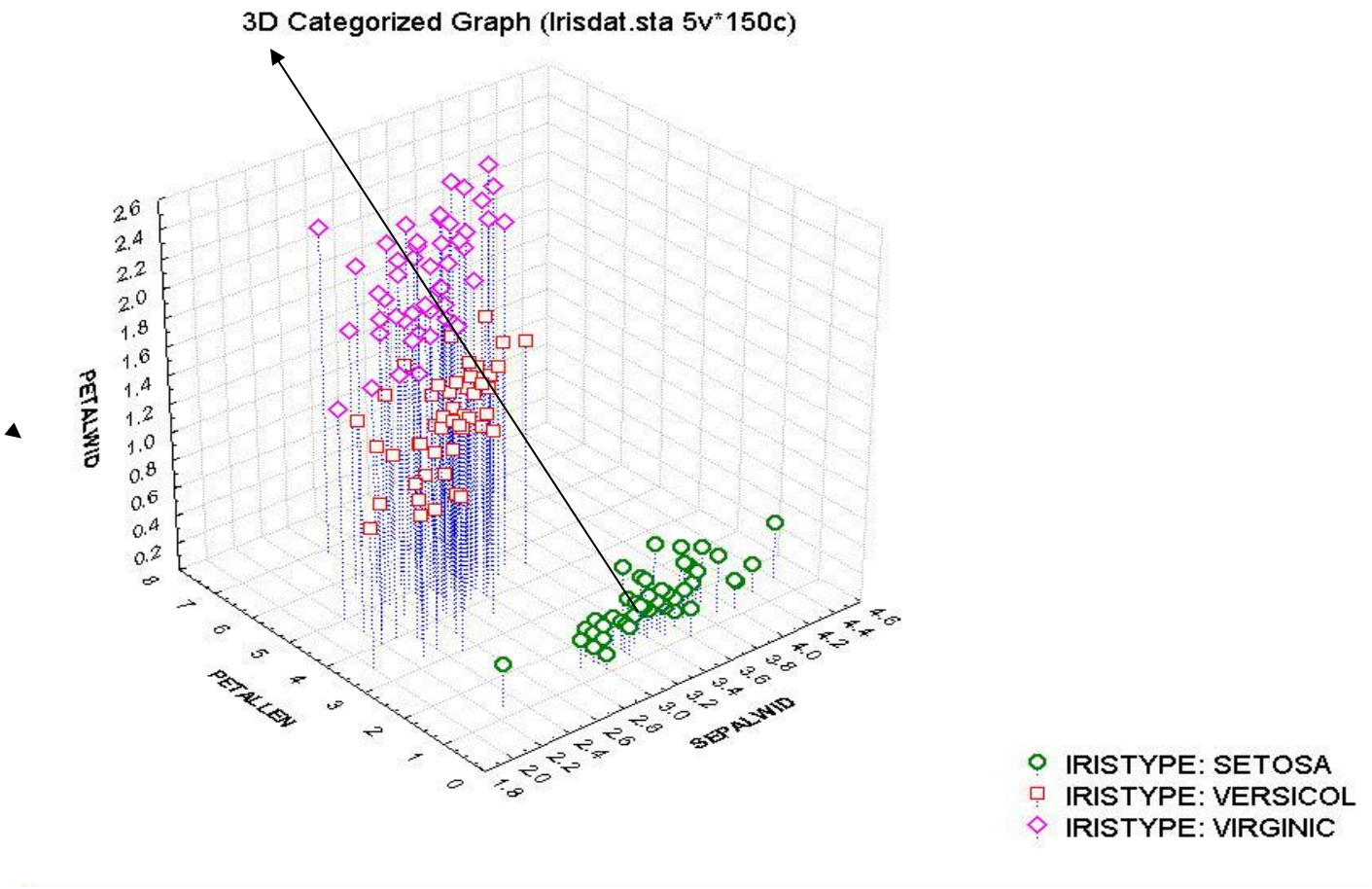
Рональд Фишер
1890 - 1962

Наиболее известным статистиком XX века, безусловно, является Рональд Фишер (1890–1962), который заложил основы дисперсионного анализа. Кроме того, он первым начал систематически рассматривать объекты и выборки в многомерном пространстве и анализировать их разнообразие и взаимное расположение. Ему принадлежит заслуга разработки многомерного обобщения дисперсионного анализа – дискриминантного анализа – как способа нахождения одномерного направления, в проекции на которое наиболее различаются выборки.

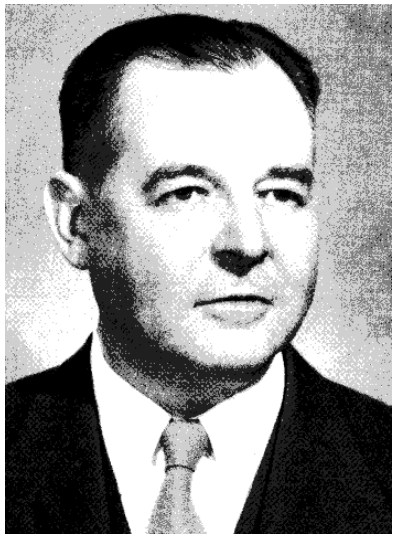
Дискриминантный анализ (MANOVA)

(однофакторный)

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7: 179-188.



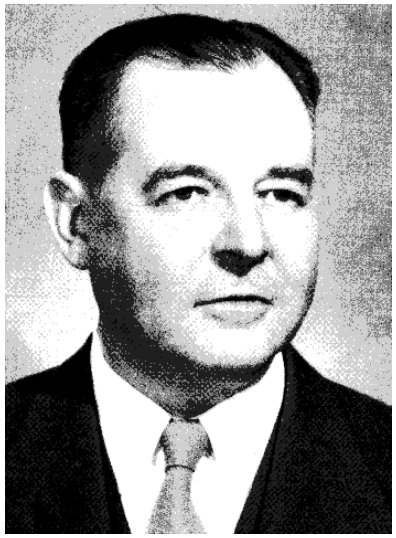
Главные компоненты



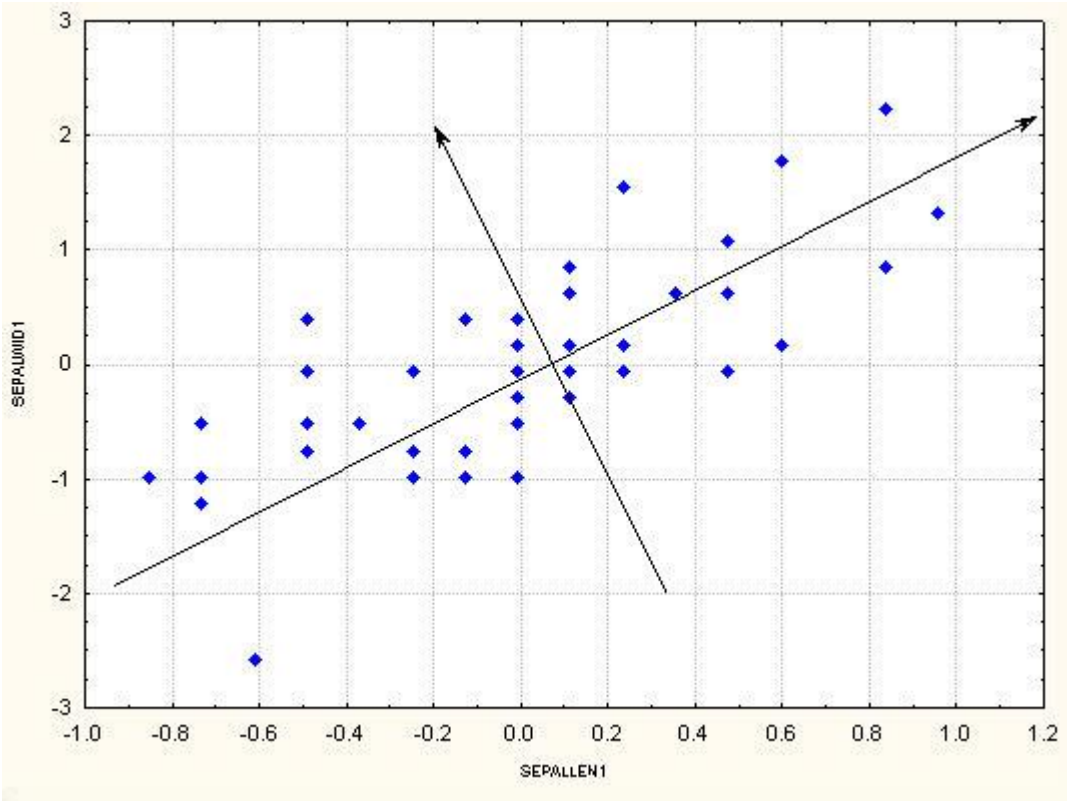
Гарольд Хотеллинг
1895-1973

Гарольд Хотеллинг (1895–1973) предложил метод главных компонент (не зная работы К.Пирсона) и канонический корреляционный анализ (Hotelling, 1933, 1936). Метод главных компонент сейчас применяется наиболее широко из всех многомерных методов, а также является базой для многих других методов. Хотеллинг был выдающимся американским экономистом, однако свою основополагающую работу по многомерному анализу опубликовал в психологическом образовательном журнале. В 1929 году шесть месяцев работал на Ротамстедской опытной станции (Великобритания) под руководством Р.Фишера.

Главные компоненты



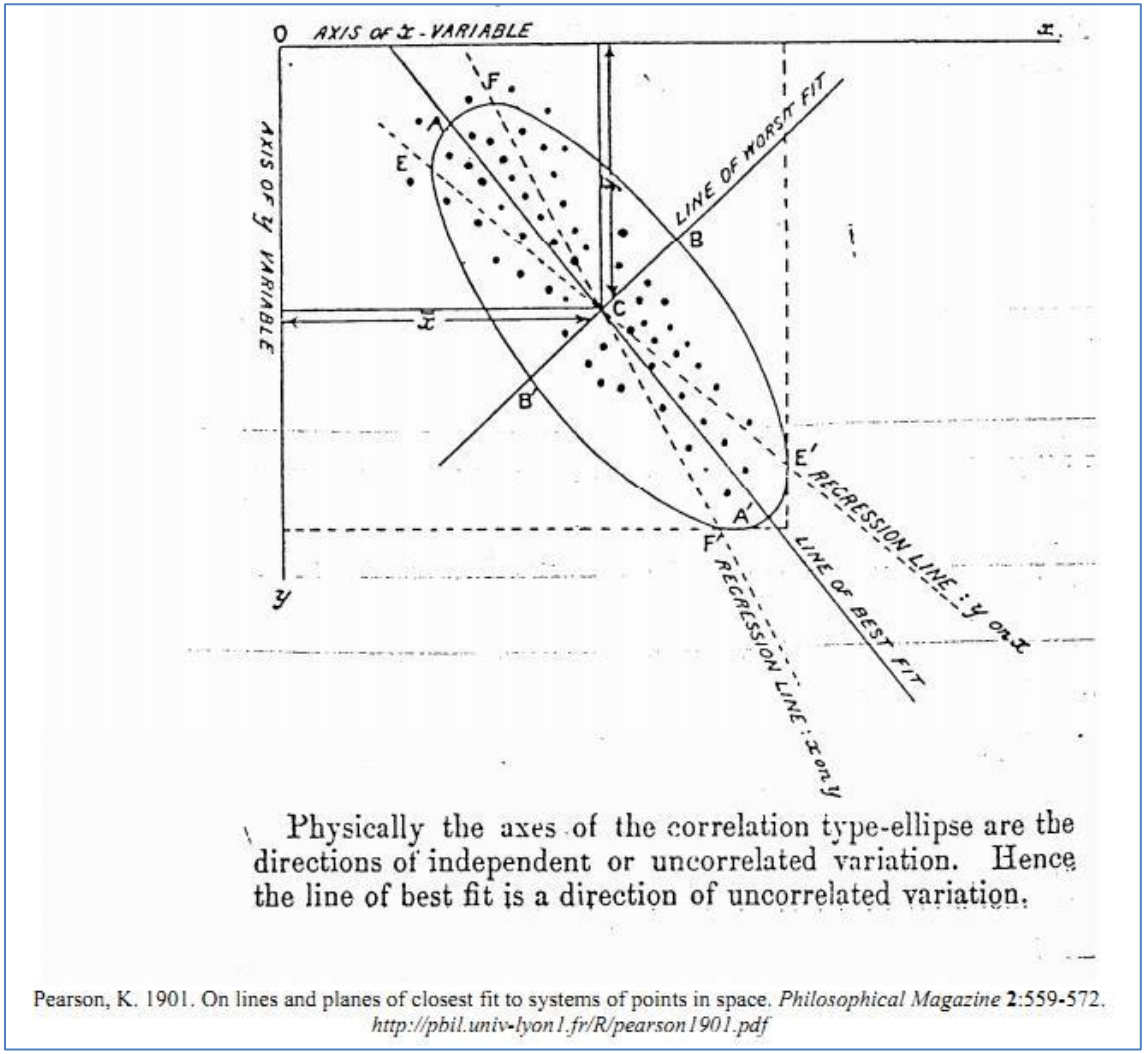
Гарольд Хотеллинг
1895-1973



Переход к главным компонентам
не меняет
расстояний между объектами

Главные компоненты (предшественники)

Карл Пирсон
1857—1936

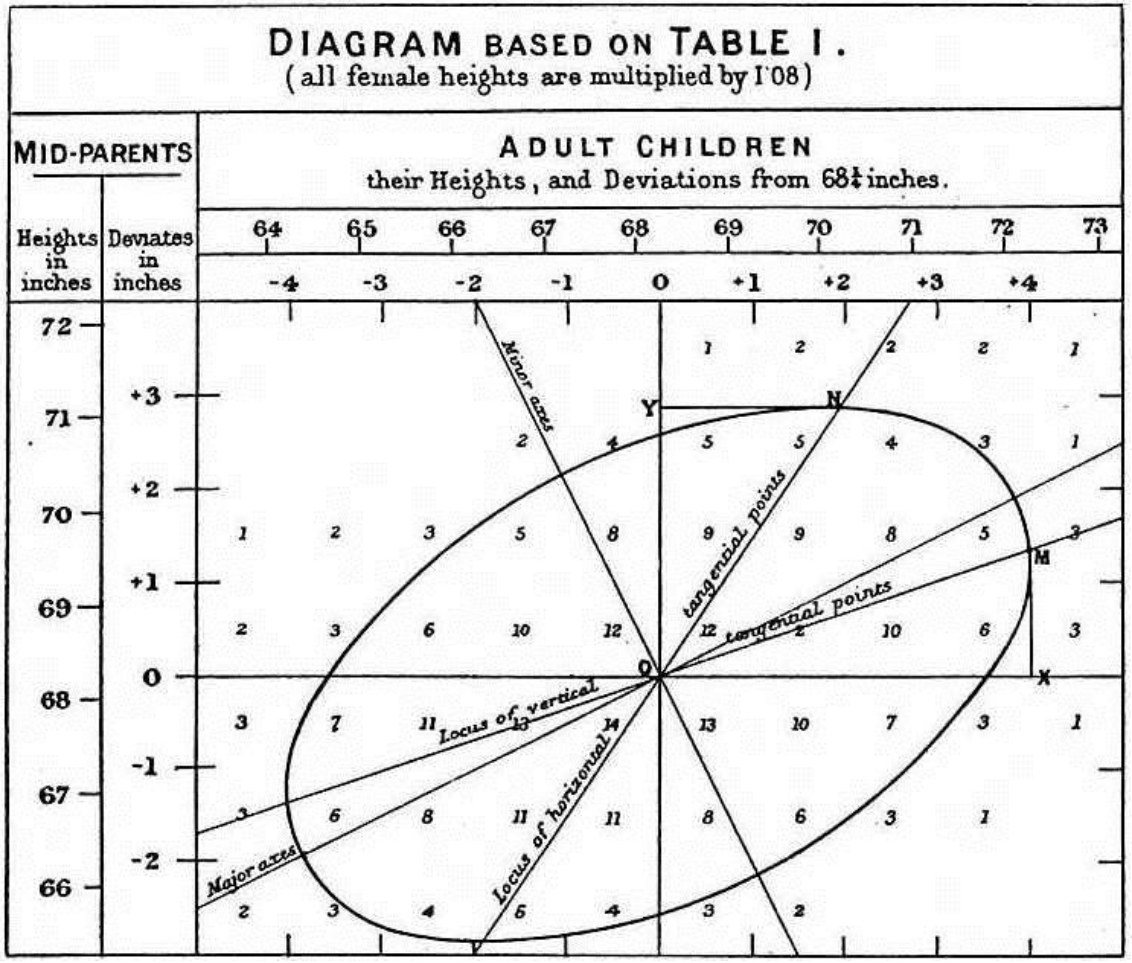


Physically the axes of the correlation type-ellipse are the directions of independent or uncorrelated variation. Hence the line of best fit is a direction of uncorrelated variation.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572. *http://pbil.univ-lyon1.fr/R/pearson1901.pdf*

Главные компоненты (предшественники)

Ф.Гальтон
1822–1911



Психометрика. Тесты

Ф.Гальтон
1822–1911

Начало научному тестированию в психометрике положил Ф.Гальтон, который пришел к необходимости измерять, кроме прочих, и психические характеристики человеческой личности: «Психометрия, необходимо твердо сказать, означает искусство охватывать измерением и числом операции ума (mind)», «Пока феномены какой-нибудь отрасли знания не будут подчинены измерению и числу, они не могут приобрести статус и достоинство науки». Ясно понимая, что человека нужно рассматривать по всей совокупности свойств как единое целое, он предложил схему обследования, в которую входили морфологические (рост, вес), физиологические (сила удара, скорость реакции) и психологические (ответы на тесты) признаки и обследовал более 9 тыс. человек. Ф.Гальтон заметил, что результаты тестов должны коррелировать друг с другом и использовал для этого коэффициент линейной корреляции.

Психометрика. Тесты

Ф.Гальтон
1822–1911

Примерно в это же время Дж.Кеттел, ученик Ф.Гальтона, предложил набор тестов, направленных именно на измерение психических свойств человека, т.е., тех, которые, с точки зрения обыденного сознания, меньше всего поддаются измерению.

Для измерения любого свойства необходима шкала, в которой можно выражать и сравнивать результаты измерений. В естественных и технических науках измерение означает сравнение с эталоном. Однако в психологии, в отличие от естественных и технических наук, очень трудно предложить какие-либо универсальные эталоны, вроде метра или килограмма. Поэтому каждый психолог разрабатывал свой собственный набор характеристик личности, а также набор тестов для их выявления. Уже в двадцатых годах прошлого века их насчитывалось больше тысячи.

Факторный анализ

Чарльз Спирмен
1863 - 1945

Ч.Спирмен предложил теорию G-фактора – генерального фактора, который через корреляции должен обнаруживаться во всех тестах и который можно интерпретировать как проявление некоей умственной энергии. Он же предложил ранговый коэффициент корреляции, носящий теперь его имя. Фактически речь шла об одномерной шкале измерения интеллектуальных способностей.

Факторный анализ

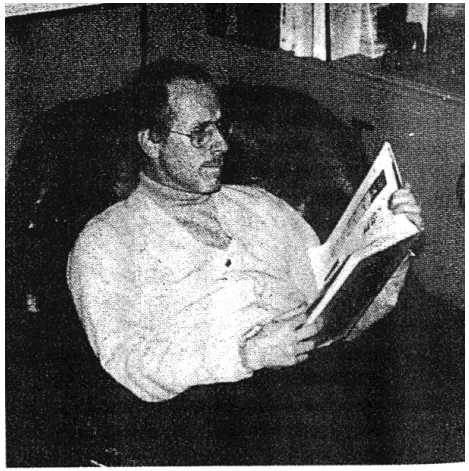
Луис Леон Терстоун
1887 - 1955

Л.Терстоун разработал свой вариант факторного анализа. В отличие от подхода Ч Спирмена, где интерпретация была определена заранее, факторный анализ Терстоуна допускал несколько групповых факторов и мог применяться к данным любой природы, а не только психологическим. ***(Следует специально отметить, что у психологов речь шла не столько о математической модели, в которой естественно рассматривать несколько факторов, а один – считать просто частным случаем, сколько о том, какой именно вариант реализуется в действительности.)***

Многомерное шкалирование

Роджер Шепард

Р.Шепард построил алгоритм неметрического шкалирования, минимизирующий различия между двумя упорядочениями: различий в исходной матрице данных и дистанций в многомерном пространстве. Особенно обнадежило то обстоятельство, что при неметрических предпосылках алгоритм практически однозначно воссоздавал метрическую структуру данных за счет избыточности числа связей между объектами.



Йозеф Крускал

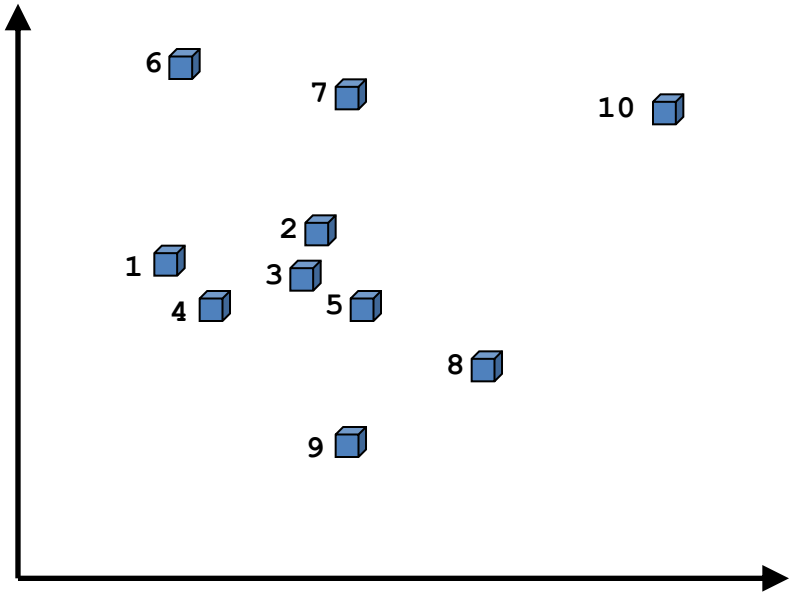
Дж.Крускал модифицировал этот алгоритм, предложив использовать квазиметрическую меру различий между двумя упорядочениями (“стресс”), сохраняющуюся при монотонных преобразованиях, и известные градиентные методы минимизации функций многих переменных. Алгоритм Крускала используется в программе STATISTICA.

Многомерное шкалирование

Матрица сходства-различия между объектами

R	1	2	3	4	5	6	7	8	9	10
1	0	52	66	64	66	60	63	67	57	64
2	52	0	63	50	55	48	56	54	45	45
3	66	63	0	51	57	57	55	54	51	64
4	64	50	51	0	48	44	54	56	50	49
5	66	55	57	48	0	33	36	25	33	54
6	60	48	57	44	33	0	37	42	40	52
7	63	56	55	54	36	37	0	39	41	61
8	67	54	54	56	25	42	39	0	27	54
9	57	45	51	50	33	40	41	27	0	41
10	64	45	64	49	54	52	61	54	41	0

Представление в виде точек евклидова пространства малой размерности



PLS-анализ

Herman Wold
1908 - 1992

Svante Wold
1941-

H. Wold. Soft modelling. The basic design and some extensions. In: Systems Under Indirect Observation (Eds. K.-G. Joreskog, H. Wold). V. I-II, North-Holland, Amsterdam, **1982**.

Svante Wold*, Axel Ruhe,* ***Herman Wold,*** *and W.J. Dunn* (1984). The collinearity problem in linear regression. the **partial least squares (PLS**) approach to generalized inverses. SIAM J. Sci. Stat. Comp. 5:735-743.

Wold S, *Hellberg S, T L, Sjostrom M,* ***Wold H.*** **PLS** Model Building: Theory and applications. **PLS** modeling with latent variables in two or more dimensions. Frankfurt am Main, **1987**.

Svante Wold, *Michael Sjostrom, Lennart Eriksson.* **PLS**-regression: a basic tool of chemometrics // Chemometrics and Intelligent Laboratory Systems 58 **2001** 109–130.

PLS-анализ

PLS – Projection to Latent Structure
(Устаревшее (хотя и часто используемое) название: PLS – Partial Least Squares)

Основная идея: вместо метода наименьших квадратов используется принцип максимальной ковариации.
На методе наименьших квадратов основана ВСЯ теория современного многомерного статистического анализа.
PLS-анализ такого теоретического обоснования НЕ имеет.
Однако на практике показывает значительно лучшие результаты.

PLS - регрессия

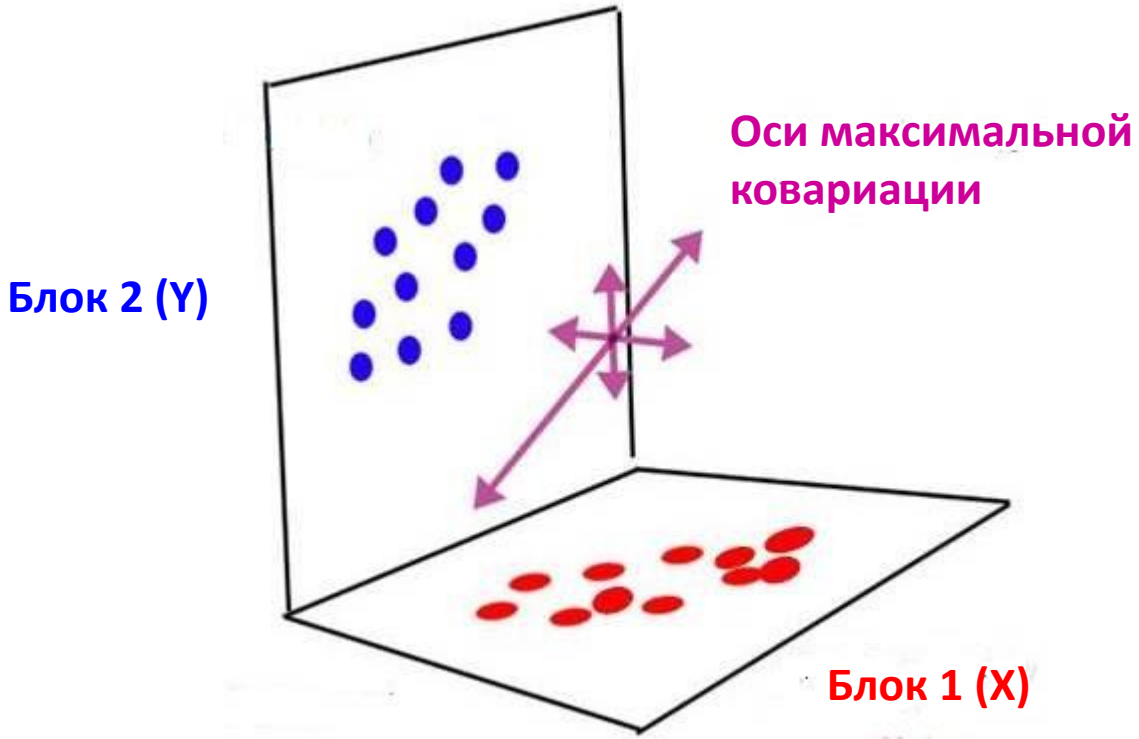
В PLS-регрессии максимизируется ковариация между зависимой переменной и аппроксимирующей ее линейной комбинацией независимых переменных.

Обязательным условием является разбиение выборки на обучающую и контрольную (проверочную) части. На обучающей части оцениваются коэффициенты аппроксимации.

На контрольной части проверяется качество предсказания.

2B-PLS анализ

В 2B-PLS-анализе максимизируется ковариация между линейными комбинациями двух совокупностей независимых переменных. Есть в программе PAST.



Этапы биостатистики

I. **Биометрия, психометрика**

Классические линейные методы

\+ кластерный анализ

II. **Математическая статистика**

Классические линейные методы

III. **Биостатистика, хемометрика**

Нелинейные методы

PLS-методы

О парадигмах биостатистики

Считается, что теоретической основой биостатистики является математическая статистика, которая, в свою очередь, исследует обратные задачи теории вероятностей. Однако математическая статистика по-настоящему освоила только классические линейные методы. Нелинейные методы (коэффициенты сходства/различия, кластерный анализ, карты Кохонена, многомерное шкалирование и т.д.) **до сих пор остаются неохваченными ввиду их явной эвристичности. Что касается PLS-методов, то здесь ситуация еще сложнее. Вся классическая математическая статистика базируется на вероятностном подходе и парадигме метода наименьших квадратов (МНК). В нелинейных методах нет вероятностного подхода, а в PLS-методах отказались от метода наименьших квадратов. Вместо него используется принцип максимальной ковариации, который *на практике* работает гораздо лучше, чем МНК, особенно на больших массивах.**

Таким образом, классическая математическая статистика явно не оправдывает своих притязаний на роль лидера для биостатистики.

О парадигмах биостатистики

Классическая математическая статистика явно не оправдывает своих притязаний на роль лидера для биостатистики.

В чем причина? На мой взгляд, их две.

Первая – переоценка роли математики и недооценка роли биологии.

Вторая – переоценка вероятностного подхода и недооценка геометрического.

Достоверность

Принятая в статистике совокупность способов убедить рецензентов (а, заодно, и себя) в том, что Ваши результаты можно публиковать ☺

Классическая практика математической статистики заключается в том, что мы идеализируем те условия, в которых были получены данные, например, предполагаем существование и многомерную нормальность распределения объектов, отсутствие систематических ошибок, бесконечно большой размер выборки и т.д. В этих идеализированных условиях мы рассчитываем вероятность случайного получения нашего результата и, если она оказывается достаточно мала, делаем вывод, что наша гипотеза статистически подтверждается.

Более современный способ – компьютерное моделирование. Требует специализированного программного обеспечения и мощных вычислительных средств. Никаких дополнительных предположений не требуется. В исходные данные намеренно вносятся искажения (например, с помощью бутстрепа) и новые данные просчитываются теми же методами, что и исходные. Операция повторяется много раз. Разброс результатов характеризует надежность полученных выводов.

Бутстреп

Б.Эфрон предложил размножать исходную выборку. Пусть она состоит из *N* элементов. Новую выборку получим следующим образом. С помощью датчика случайных чисел с равными вероятностями выберем любой элемент исходной и включим его копию в новую выборку. Повторим процесс *N* раз. Выборка сформирована.

Б. Эфрон. Нетрадиционные методы многомерного статистического анализа. –М.: ФиС, 1988. 263 с.

На основе только исходной выборки, мы всегда можем получить бутстреп-модели генеральных распределений и для нулевой и для альтернативной гипотезы, и после этого вычислить ошибки первого и второго рода для любого выбранного нами порогового значения.

О применимости t-критерия Стьюдента

Имеются две выборки и один признак. Требуется оценить достоверность различий средних.

Обычная рекомендация отечественной биометрии заключается в том, что надо применить t-критерий Стьюдента:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

При этом обращается внимание на то, что, строго говоря, этот критерий выведен при двух условиях, которые надо проверять дополнительно:

1) **нормальность распределения признака в обеих сравниваемых группах;**
2) **равенство генеральных дисперсий двух сравниваемых групп.**

Из условия 1) вытекает еще одно условие:

3) признак должен быть количественным,

иначе ни о какой нормальности распределения говорить не приходится.

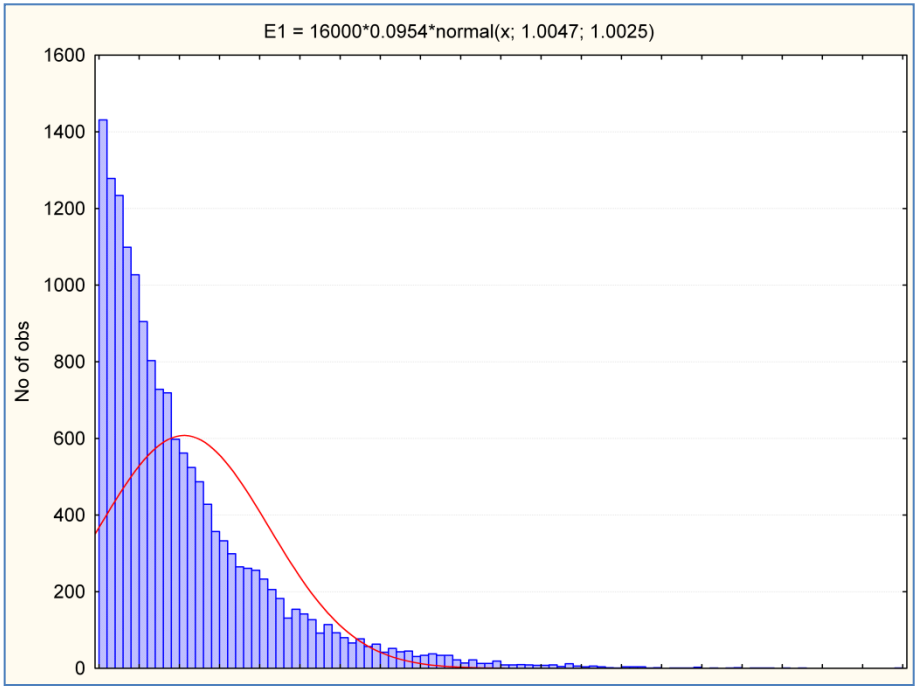
Что делать, если хотя бы оно из этих условий не выполняется? Предлагается использовать другие методы, например, непараметрические типа Манна-Уитни.

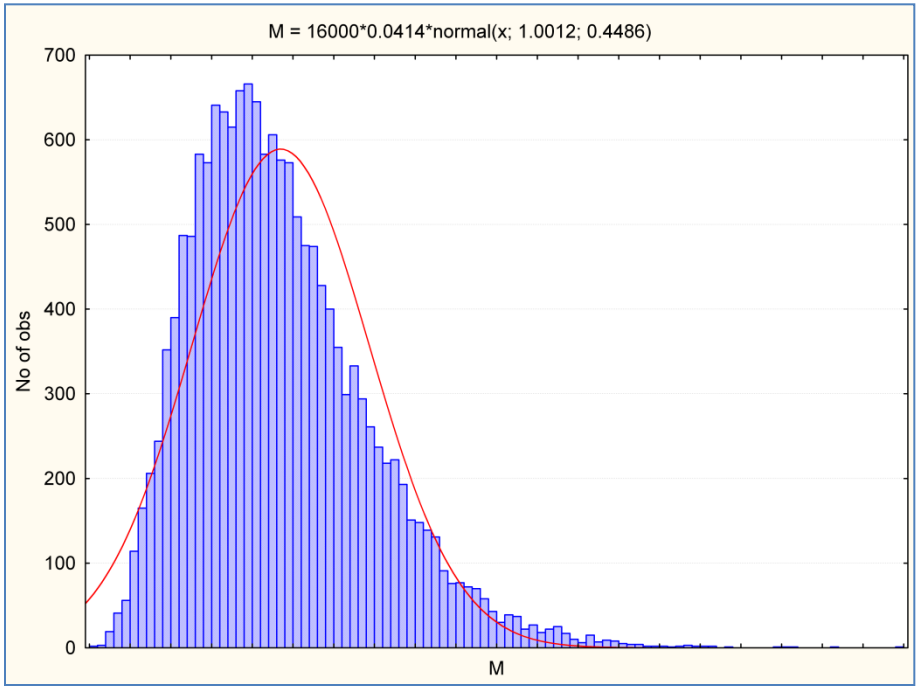
О применимости t-критерия Стьюдента

Я считаю, что в этой рекомендации содержится, как минимум, пять неточностей:

1. **Стьюдент – это не Стьюдент (настоящая фамилия – Госсет).**
2. **Критерий Стьюдента на предыдущем слайде – это не критерий Стьюдента, а критерий Уэлша. Именно он обычно используется в отечественной биометрии под названием критерия Стьюдента.**
3. **В отличие от критерия Стьюдента, критерий Уэлша не требует равенства генеральных дисперсий (условие 2).**
4. **В обоих критериях речь идет о средних. А средние при любых распределениях исходных данных всегда распределены приближенно нормально в силу центральной предельной теоремы. Поэтому *на практике* t-критерий Уэлша будет работать *всегда*. Поэтому нормальность (условие 1) проверять не обязательно.**
5. **По тем же причинам, вопреки условию 3, критерий Уэлша можно применять и для других типов признаков, например, ранговых и двоичных. Средние по ним тоже всегда будут распределены приближенно нормально.**

О применимости t-критерия Стьюдента





Слева: экспоненциальное распределение(λ=1). Справа: распределение выборочных средних при n=10. Красная линия: аппроксимация нормальным распределением.

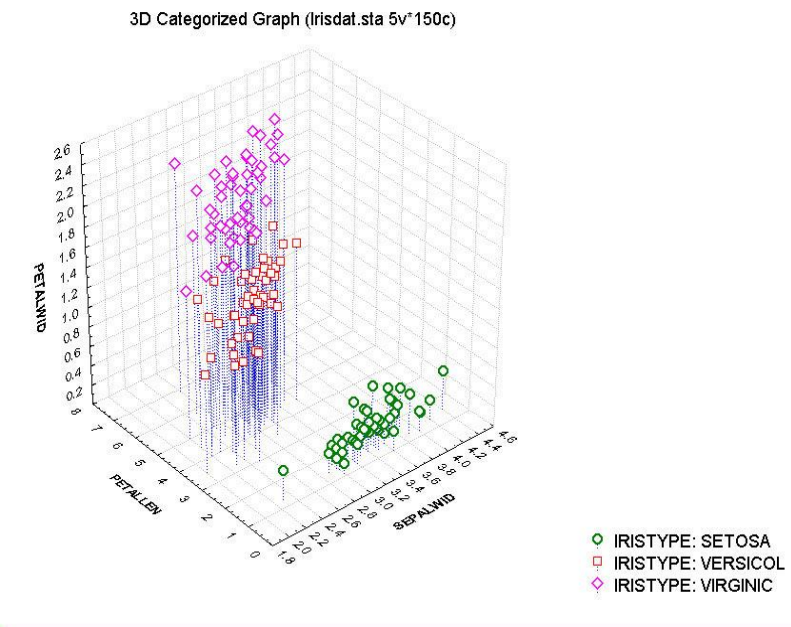
О применимости t-критерия Стьюдента

Как правило, во всех случаях вместо t-критерия Стьюдента и вместо непараметрического метода Манна–Уитни лучше использовать t-тест Уэлша, который не требует однородности дисперсий (Ruxton, 2006). **Вы всегда получите биологически осмысленные результаты.**

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U-test. *Behavioral Ecology*, *17*(4), 688-690.

Вывод: биологические задачи надо решать на биологическом уровне строгости

Геометрический подход



Что такое геометрический подход?

Наши органы чувств приспособлены к восприятию расстояний и взаимного расположения предметов в трехмерном пространстве. Поэтому евклидова геометрия оказалась первой наукой, которую создало человечество для понимания свойств окружающего мира более двух тысяч лет назад. В середине XIX века стало ясно, что геометрические отношения можно трактовать гораздо шире. Во-первых, можно рассматривать пространства с размерностью больше трех, вплоть до бесконечности. Во-вторых, пространственноподобные отношения, т.е. отношения расстояния и взаимного расположения, можно искать и исследовать везде, где есть отношения сходства/различия между объектами любой природы. Чаще всего объектами являются особи. Однако в качестве объектов можно рассматривать как популяции, сообщества, их состояния, динамику, поведение и другие характеристики, так и органы, ткани, гены, профили их экспрессии и т.д. Выбор объекта определяется задачей исследования. Однако специфика объекта не играет особой роли, методы анализа, как правило, одни и те же.

Геометрический подход

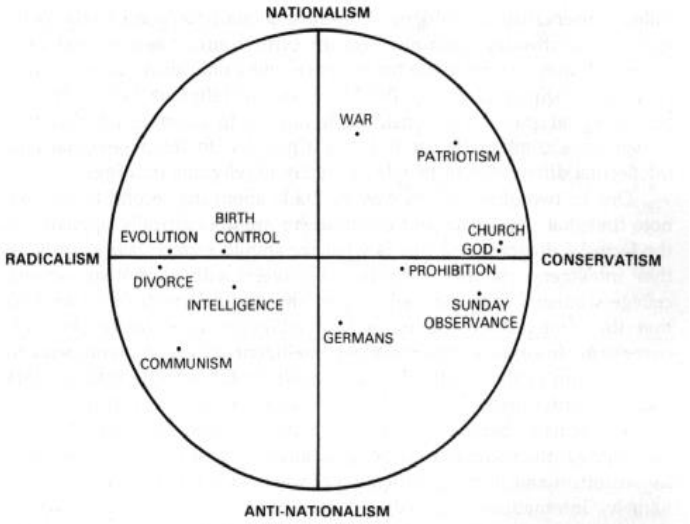


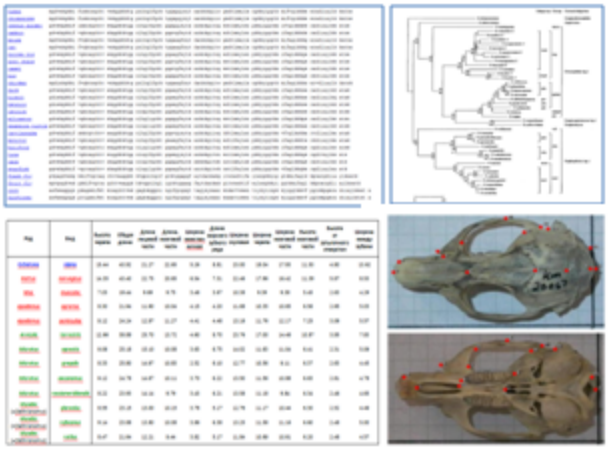
Figure 4. Factor study of radicalism with attitude scales by Thelma Gwinn Thurstone.

Таким образом, суть геометрического подхода заключается в том, что в качестве модели наших объектов мы выбираем совокупность точек в некотором многомерном пространстве, а отношения сходства/различия между объектами отображаем расстояниями между точками в этом пространстве.

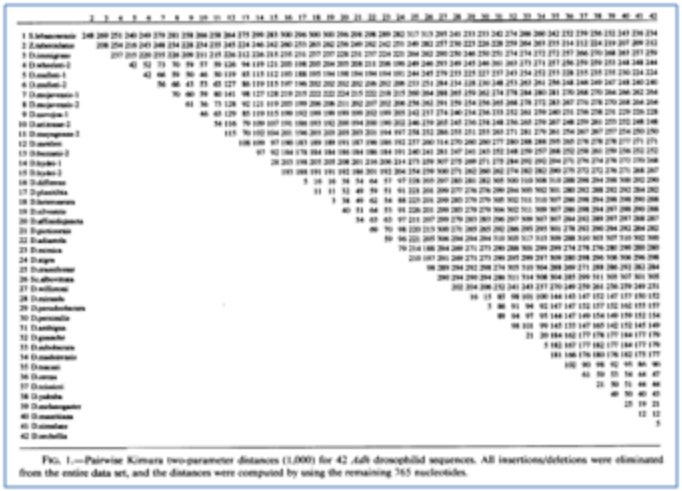
Если предполагается, что объекты извлечены из генеральной совокупности, распределенной в этом же пространстве в соответствии с некоторым вероятностным законом (это тоже модель!), то мы имеем дело с многомерным статистическим анализом (метод главных компонент (факторный анализ), множественная регрессия, канонический анализ, дискриминантный анализ).

Это предположение не является обязательным. Существуют и успешно применяются в биологии методы, которые его не используют: кластерный анализ, многомерное шкалирование, самоорганизующиеся карты признаков, нейронные сети и т.д. **Но и в классическом многомерном статистическом анализе необходимо разделять геометрическую и вероятностную составляющие.**

Представление объектов набором точек



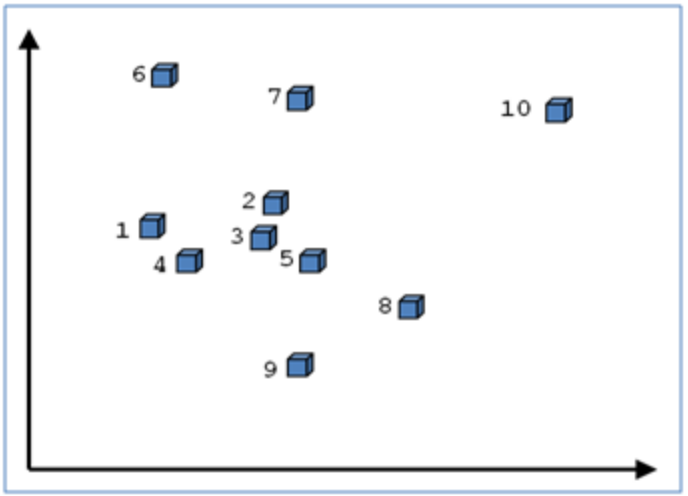
Описание любого типа



Матрица любых коэффициентов сходства-различия между объектами

N_Sp	1	2	3	4	5	6	7	8	9
1_S.lebaaonensis	0.00	15.75	16.40	15.84	15.49	15.78	16.70	16.76	16.06
2_Z.tuberculatus	15.75	0.00	14.42	15.94	14.70	15.59	15.75	15.94	15.10
3_D.immigrans	16.40	14.42	0.00	15.39	14.66	14.83	15.33	15.03	14.46
4_D.wheelcri-2	15.84	15.94	15.39	0.00	6.48	7.21	8.54	8.37	7.68
5_D.mulleri-1	15.49	14.70	14.66	6.48	0.00	6.48	8.12	7.68	7.07
6_D.mulleri-2	15.78	15.59	14.83	7.21	6.48	0.00	7.48	8.12	6.56
7_D.mojavensis-1	16.70	15.75	15.33	8.54	8.12	7.48	0.00	8.37	7.75
8_D.mojavensis-2	16.76	15.94	15.03	8.37	7.68	8.12	8.37	0.00	7.81
9_D.aavojoa-1	16.06	15.10	14.46	7.68	7.07	6.56	7.75	7.81	0.00

Матрица евклидовых расстояний между объектами



Набор точек в евклидовом пространстве

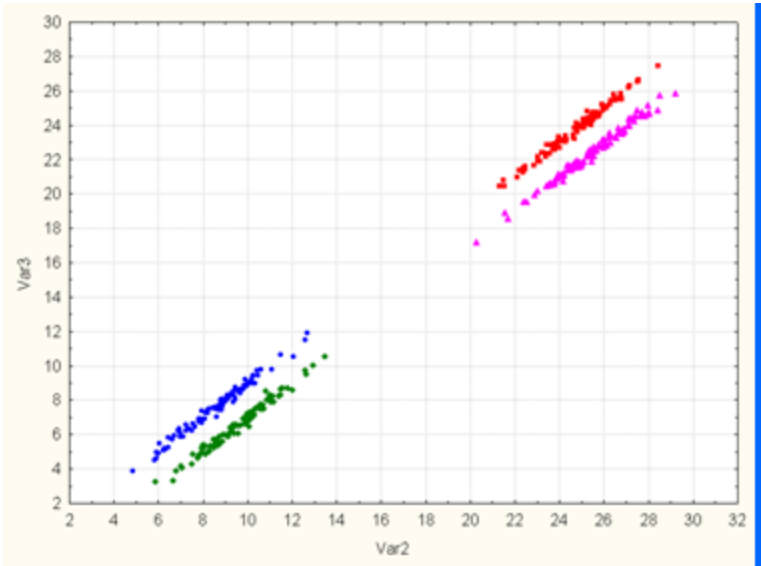
Геометрический подход

С точки зрения геометрического подхода метод наименьших квадратов ущербен. Если к одномерным методам претензий нет, там МНК работает нормально, то в многомерных методах, за исключением метода главных компонент (и факторного анализа), МНК автоматически приводит к искажению расстояний между объектами в многомерном пространстве. В частности, это напрямую касается множественной регрессии и дискриминантного анализа. Методу главных компонент повезло больше, он по построению удовлетворяет обеим парадигмам сразу, и МНК, и PLS. PLS не искажает расстояний между объектами.

Почему надо сохранять расстояния? Из биологических соображений. На предварительном этапе статистической обработки приходится выбирать признаки, нормировать, подбирать меры сходства/различия для того, чтобы адекватно отразить разнообразие биологических объектов расстояниями между представляющими их точками в многомерном пространстве. Если расстояния не сохраняются, искажение содержательных результатов неизбежно.

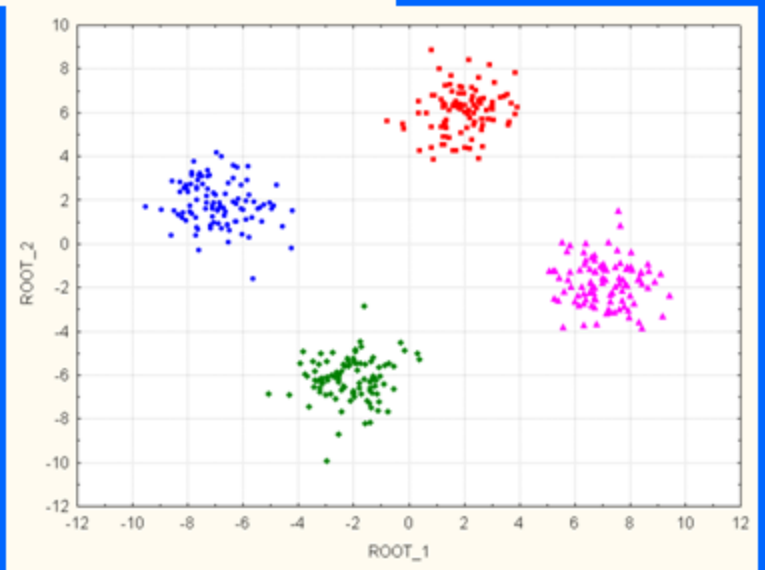
Геометрический подход

Дискриминантный анализ



Максимизация функционала Фишера

D=16



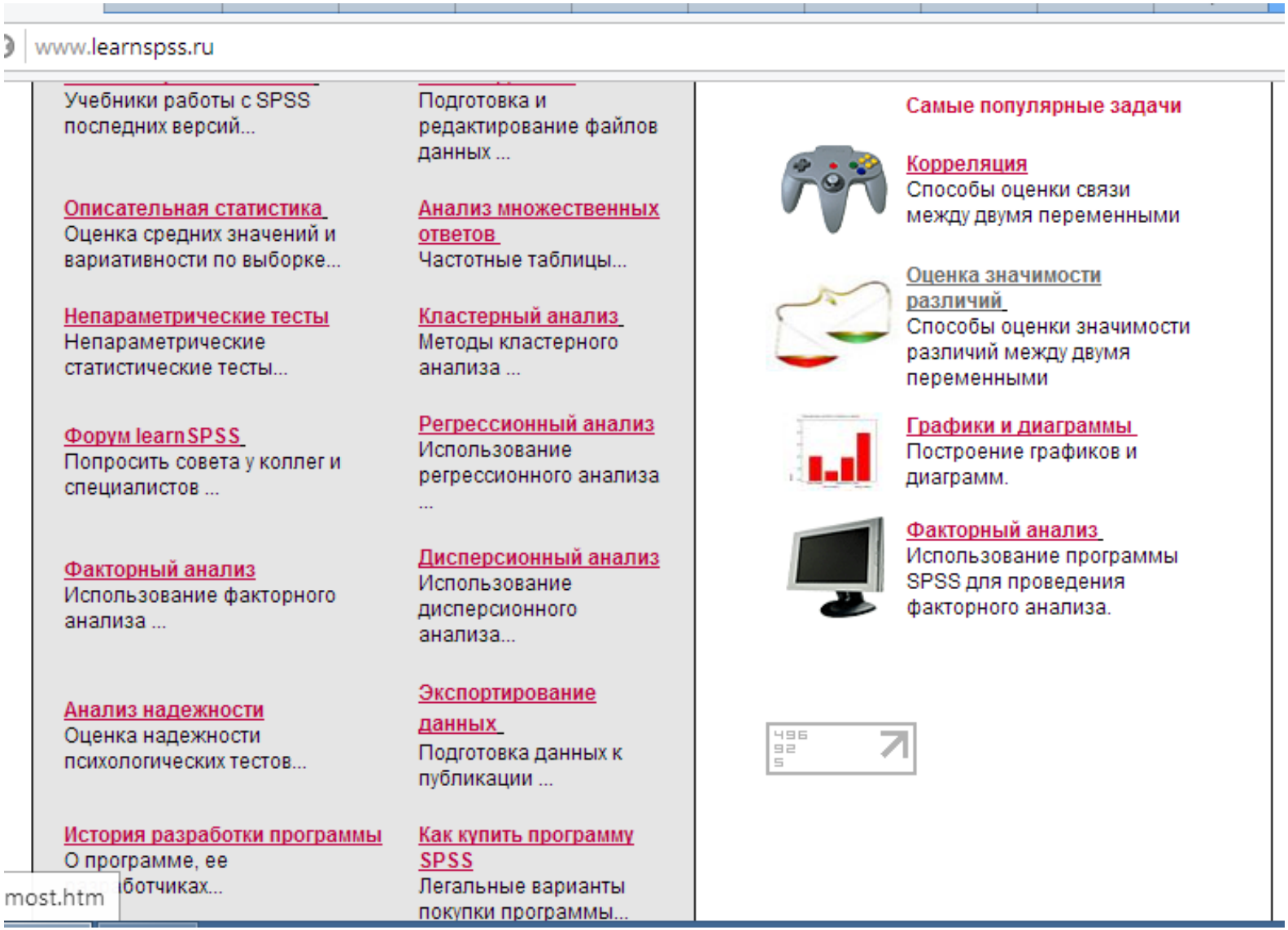
Статистические пакеты

На сегодняшний день в мире насчитывается больше **1000** статистических пакетов. Статистические пакеты делятся на несколько категорий. Во-первых, коммерческие (платные) и находящиеся в свободном доступе (бесплатные). Во-вторых, универсальные и специализированные. В-третьих, интерактивные и предлагающие пакетную обработку. В-четвертых, закрытые и настраиваемые.

Наиболее известны зарубежные статистические пакеты: **SAS, SPSS, STATISTICA, R, S-plus** и т.п. Для биологов очень полезен **PAST.**

Необходимо отметить, что существует минимальный набор статистических методов анализа, который включен во все распространенные пакеты:
описательная статистика (базовые статистические методы);
дисперсионный анализ; дискриминантный анализ;
непараметрическая статистика;
контроль качества; анализ выживаемости;
кластерный анализ;
факторный анализ;
регрессионный анализ;
обработка данных (сортировка, отбор, *трансформация данных*).

Пакет SPSS


www.learnspss.ru
Учебники работы с SPSS последних версий...
Подготовка и редактирование файлов данных ...
Описательная статистика
Оценка средних значений и вариативности по выборке...
Анализ множественных ответов
Частотные таблицы...
Непараметрические тесты
Непараметрические статистические тесты...
Кластерный анализ
Методы кластерного анализа ...
Форум learnSPSS
Попросить совета у коллег и специалистов ...
Регрессионный анализ
Использование регрессионного анализа ...
Факторный анализ
Использование факторного анализа ...
Дисперсионный анализ
Использование дисперсионного анализа...
Анализ надежности
Оценка надежности психологических тестов...
Экспортирование данных
Подготовка данных к публикации ...
История разработки программы
О программе, ее
ботчиках...
Как купить программу SPSS
Легальные варианты покупки программы...
most.htm
Самые популярные задачи
Корреляция
Способы оценки связи между двумя переменными
Оценка значимости различий
Способы оценки значимости различий между двумя переменными
Графики и диаграммы
Построение графиков и диаграмм.
Факторный анализ
Использование программы SPSS для проведения факторного анализа.

Пакет Statistica

StatSoft®

Мы поставили и развили в России и мире
современные технологии анализа данных!

Пресс-центр Мероприятия Контакты

Главная О компании **Продукты** Решения Обучение Консалтинг Ресурсы Портал ТВ

Поиск

STATISTICA Base

Продукт предоставляет широкий набор основных статистик в понятном интерфейсе со всеми преимуществами, простотой и мощностью технологий *STATISTICA*.

STATISTICA Base включает все графические инструменты *STATISTICA*, а также следующие процедуры:

- Описательные и внутригрупповые статистики, разведочный анализ данных
- Корреляции
- Быстрые основные статистики и блоковые статистики
- Интерактивный вероятностный калькулятор
- Т-критерии (и другие критерии групповых различий)
- Таблицы частот, сопряженности, флагов и заголовков, анализ многомерных откликов
- Множественная регрессия
- Непараметрические статистики

Есть вопросы?
Специалисты StatSoft
всегда на связи.

Продукты

Общий обзор
- *STATISTICA Base*

STATISTICA Advanced
Промышленная

Пакет Statistica

Пакет Statistica хорошо сбалансирован по соотношению "мощность/удобство". Наличие достаточно широкого спектра функциональных алгоритмов делает его достаточно привлекательным для статистиков-профессионалов.

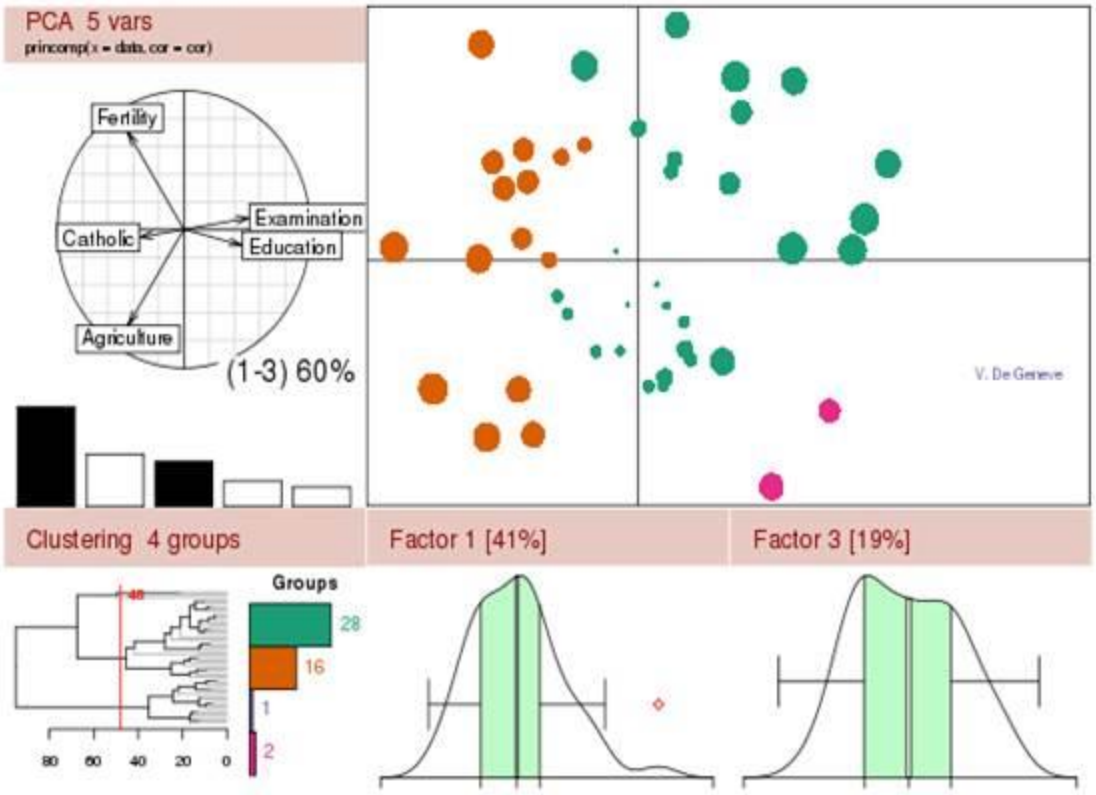
Средства манипулирования исходными данными в пакете Statistica хорошо развиты. Зачастую для проведения определенного вида анализа требуется несколько щелчков мышью. Сильной стороной пакета является графика и средства редактирования графических материалов.

http://www.statsoft.ru http://www.statsoft.com

Пакет R

www.r-project.org

The R Project for Statistical Computing



About R
What is R?
Contributors
Screenshots
What's new?

Download, Packages
CRAN

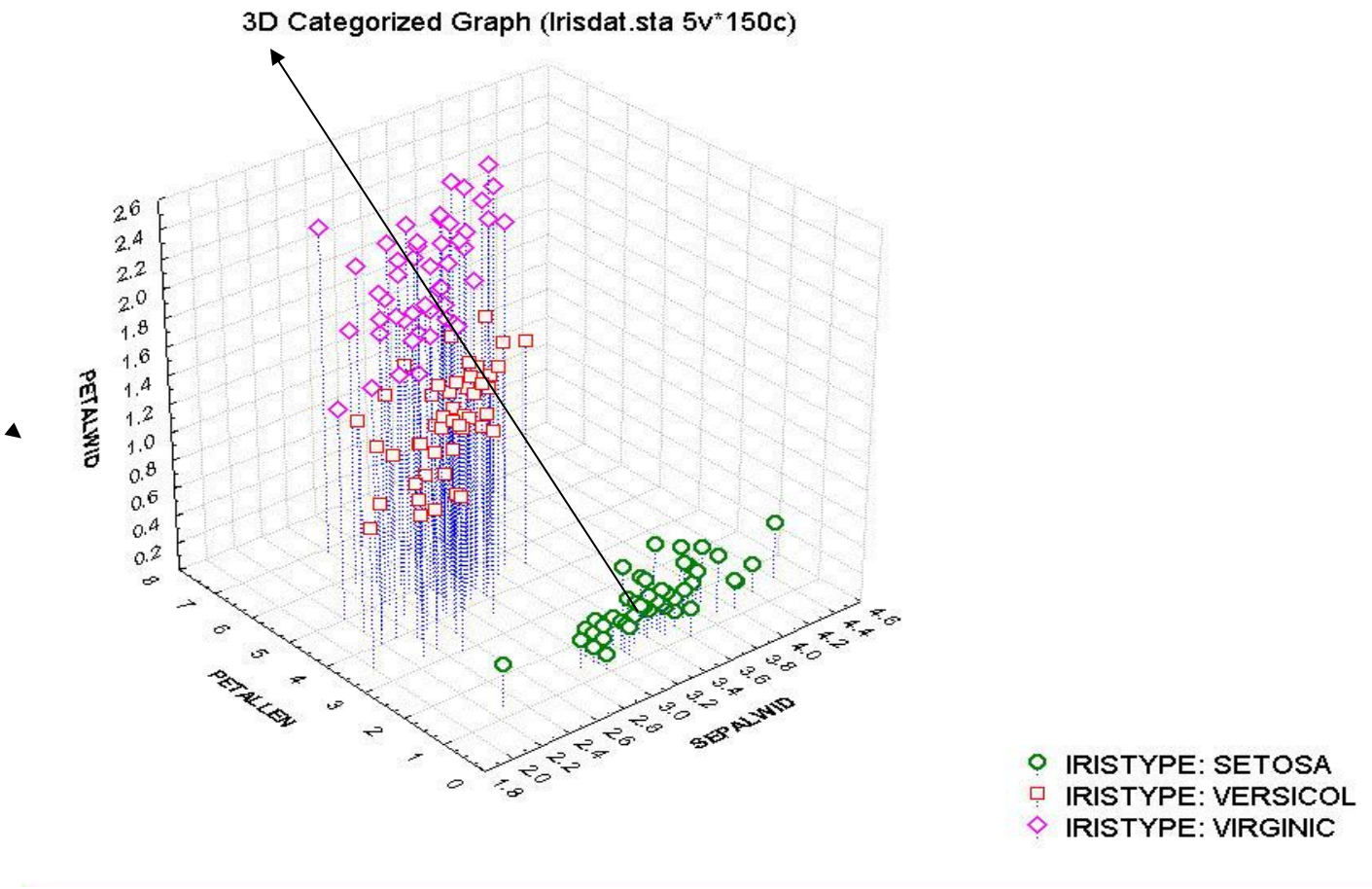
R Project
Foundation
Members & Donors
Mailing Lists
Bug Tracking
Developer Page
Conferences
Search

Getting Started:

Дискриминантный анализ (MANOVA)

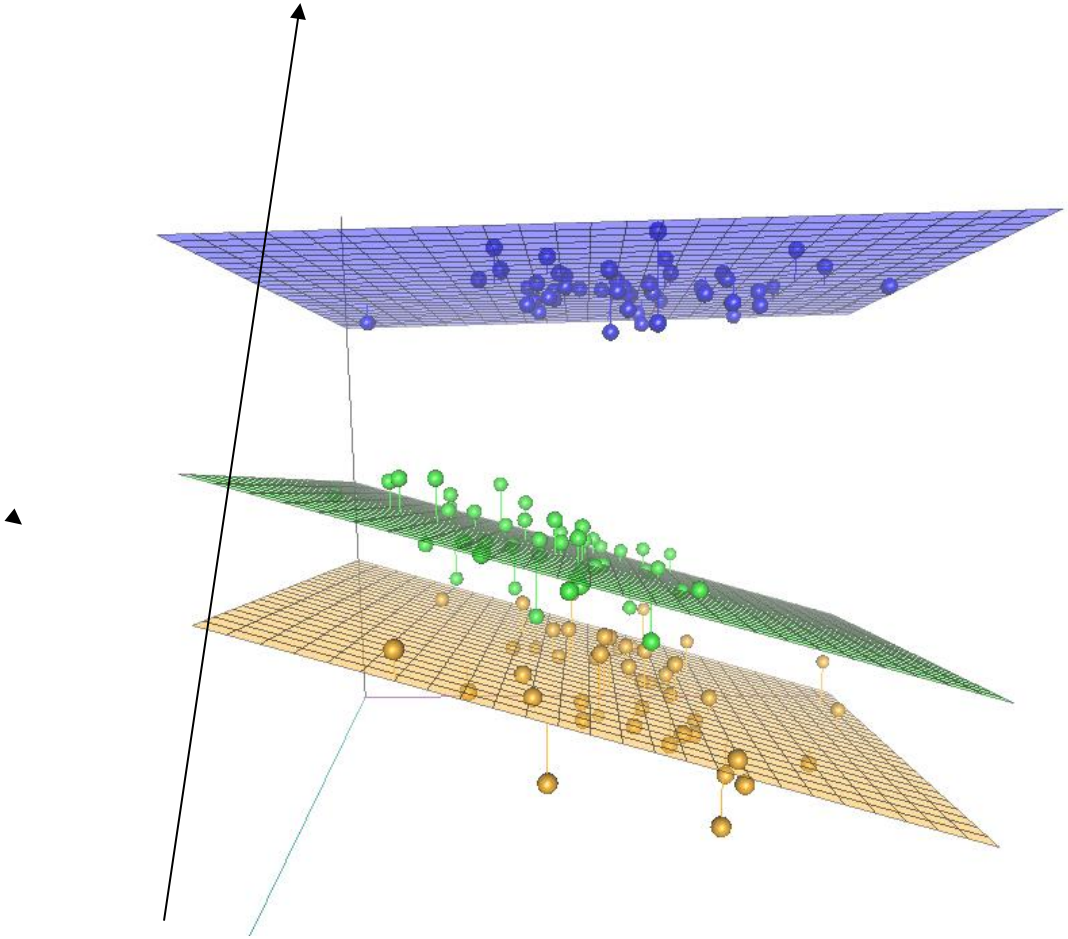
(однофакторный)

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7: 179-188.

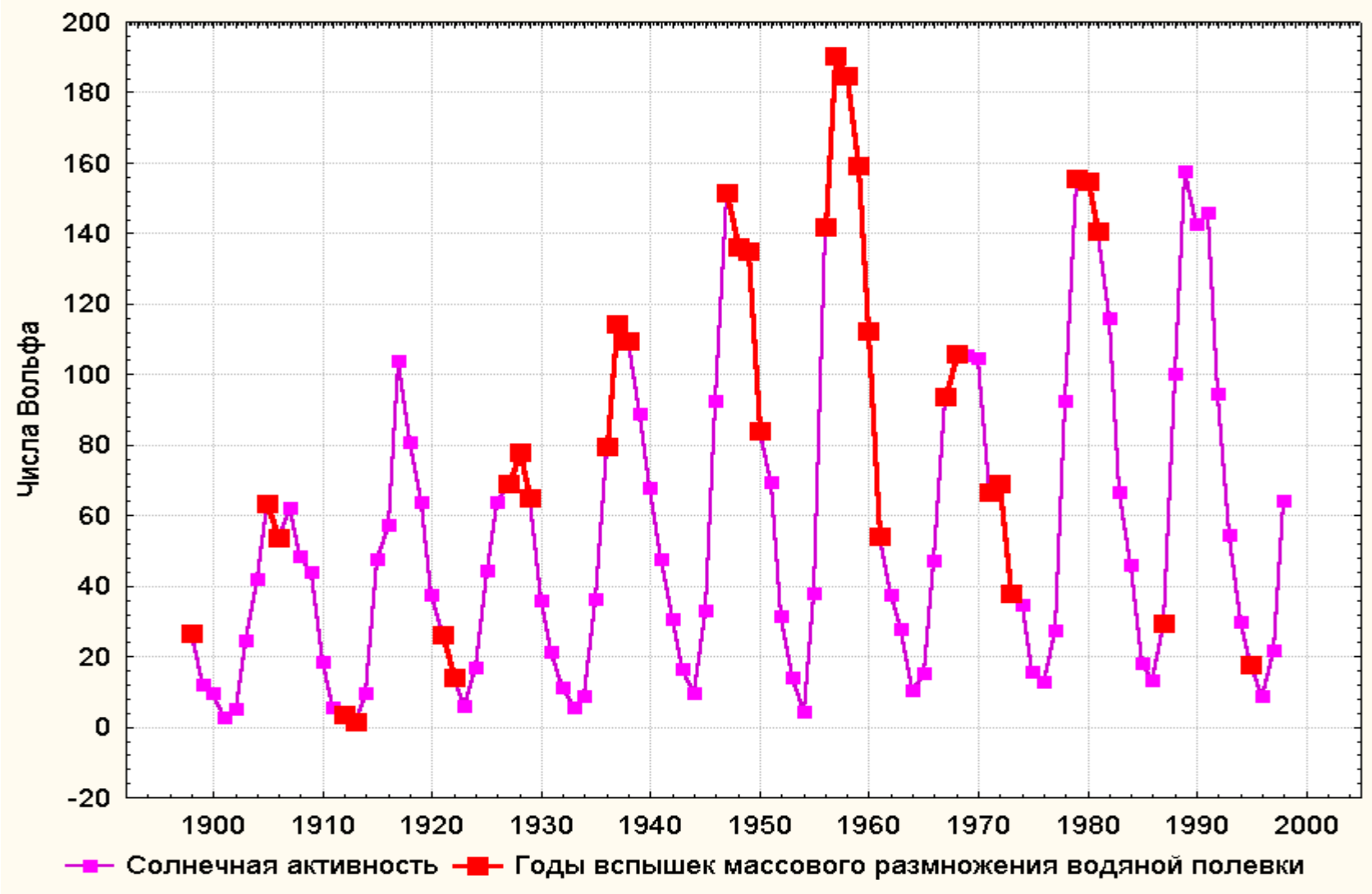


Дискриминантный анализ (MANOVA)
(однофакторный)

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7: 179-188.

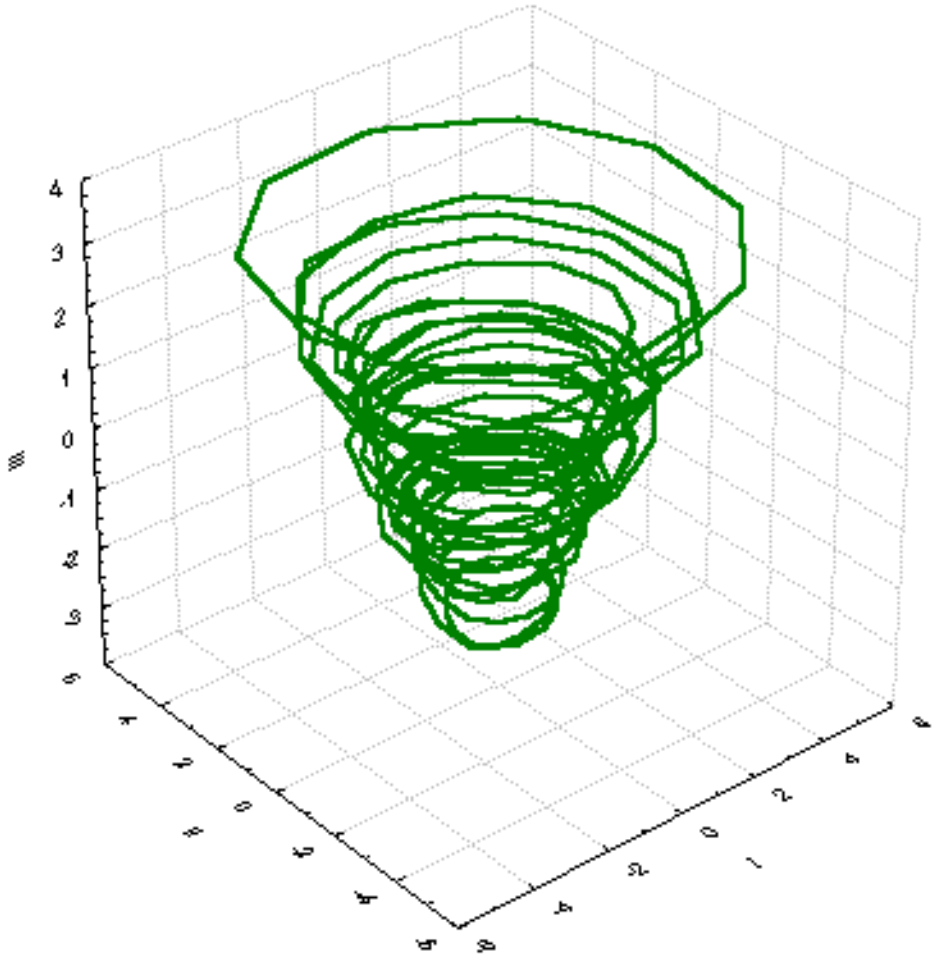


Динамика солнечной активности (числа Вольфа)



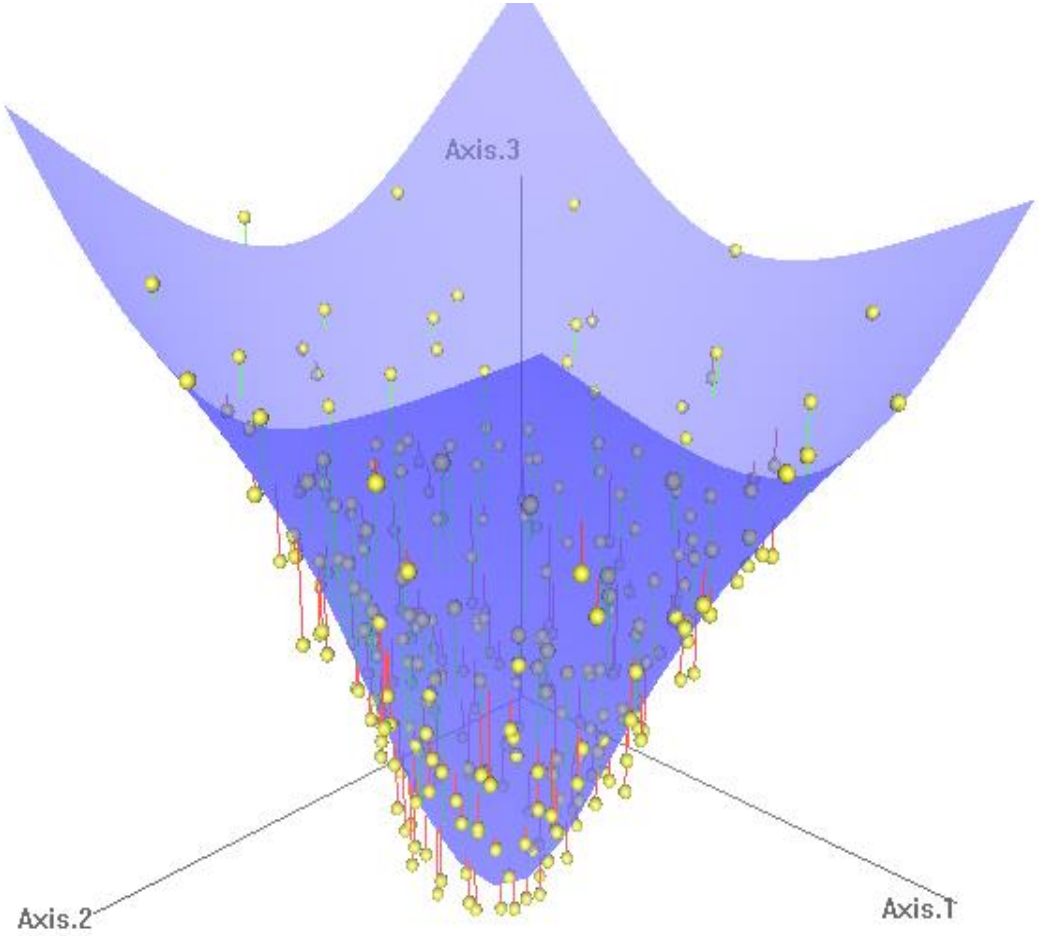
Фазовый портрет ряда Вольфа

в пространстве первых трех компонент



Фазовый портрет ряда Вольфа

Пакет R



Пакет Jacobi-4

Ефимов В.М., Полунин Д.А., Штайгер И.А.

Универсальность, модульность, простота пользования.

Входной язык приближен к естественному.

Данные и скрипты готовятся средствами Excel в формате *.csv. Пакет позволяет собирать различные сценарии обработки.

Для продвинутых пользователей реализованы циклы и подпрограммы.

Графики нет. Интерактивного режима нет.

На разработку пакета получен грант РФФИ.

Фрагмент скрипта на входном языке Jacobi-4

```
НАЧАЛО                                                          //

copyRows              Samples.csv          _cn_Samples.csv      //  Выборка строк таблицы

ЦИКЛ ПО СПИСКУ        index                GSE-96A
log                   <<index>>.csv        <<index>>-log.csv    //  Логарифмирование
copyColumns           <<index>>-log.csv    <<index>>-log.csv    //  Выборка столбцов таблицы
quantileNormalization <<index>>-log.csv    <<index>>-log.csv    //  Квантильное выравнивание
                                                                    (quantile normalization)
centre                <<index>>-log.csv    <<index>>-log.csv    //  Центрирование
normalize             <<index>>-log.csv    <<index>>-log.csv    //  Нормирование на сигму
transpose             <<index>>-log.csv    <<index>>-t.csv      //  Транспонировать
euclidean_metric      <<index>>-t.csv      <<index>>-dist.csv   //  Вычислить евклидовы расстояния
                                                                    между строками
pco                   <<index>>-dist.csv   <<index>>-pco.csv    //  Главные координаты
copyColumns           <<index>>-pco.csv    <<index>>-pco_5.csv  //  Выборка столбцов таблицы
appendRight           <<index>>-pco_5.csv  _cn_Samples.csv      //  Дописать таблицу справа
КОНЕЦ ЦИКЛА

2B-PLS                GSE-96A-pco_5.csv    GSE-97B-pco_5.csv    //  2B-PLS
copyColumns           z1.csv               z1.csv               //  Выборка столбцов таблицы
copyColumns           z2.csv               z2.csv               //  Выборка столбцов таблицы
appendRight           z1.csv               z2.csv               //  Дописать таблицу справа
euclidean_metric      z12.csv              _euclid_z12.csv      //  Вычислить евклидовы расстояния
                                                                    между строками
pco                   _euclid_z12.csv      _pco_z12.csv         //  Главные координаты
copyColumns           _pco_z12.csv         _pco_z12.csv         //  Выборка столбцов таблицы
copyColumns           _cn_Samples.csv      _cn_smp.csv          //  Выборка столбцов таблицы
appendRight           _pco_z12.csv         z12.csv              //  Дописать таблицу справа
```

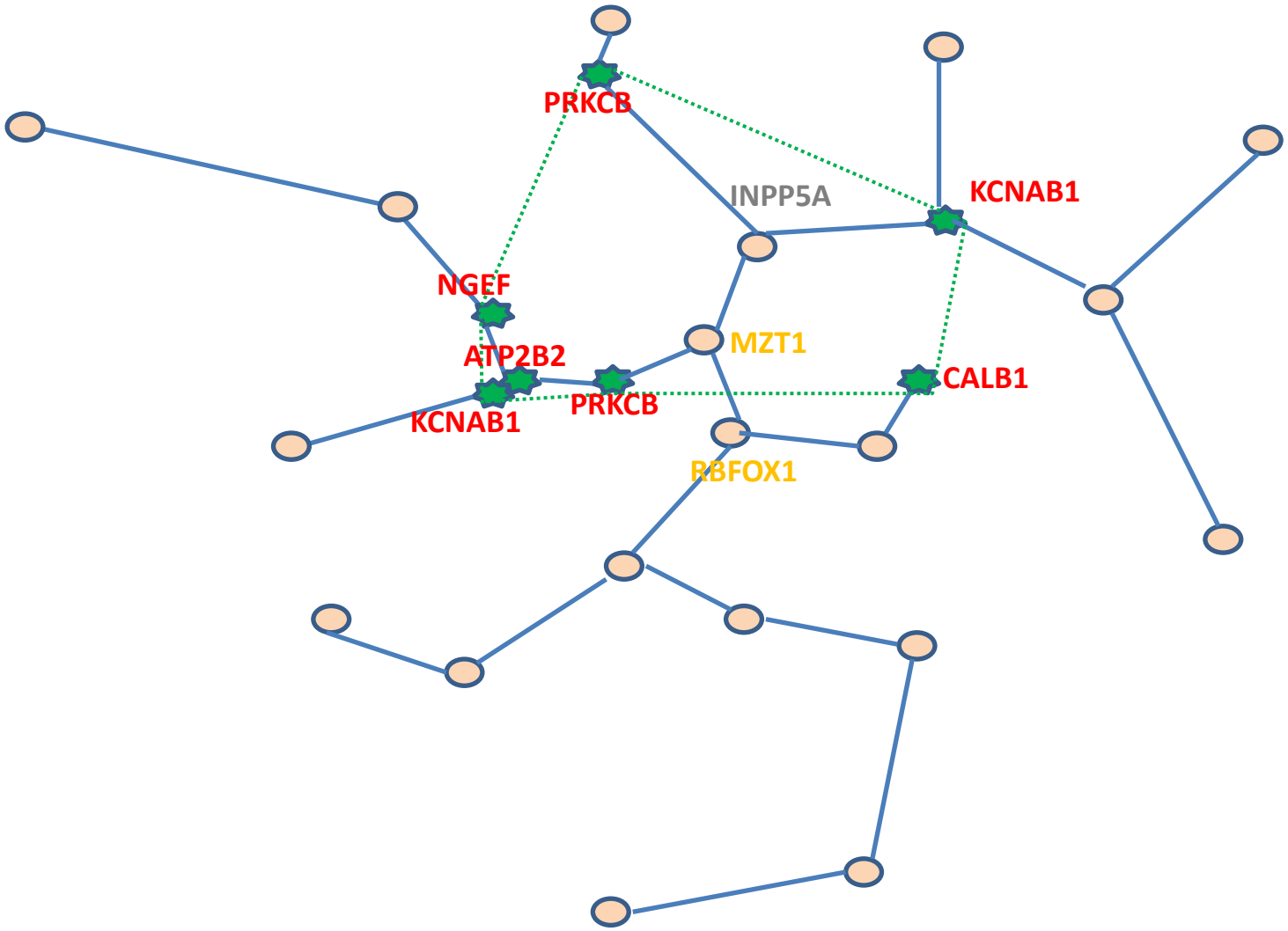
Модули

PCA (Метод главных компонент)
PCO (Метод главных координат)
SVD (Сингулярное разложение матрицы)
LDA (Линейный дискриминантный анализ)
MLR (Множественная линейная регрессия)
NMDS (Неметрическое многомерное шкалирование)
2B-PLS
PLS регрессия
Нормирование на длину
Нормирование на сигму
Нормирование на сумму
Нормирование на сумму квадратов
Центрирование
Ортогонализация по модифицированной схеме Грама–Шмидта
Логарифмирование
NNBP (нейронные сети: алгоритм обратного распространения ошибки)
convertGroupVectorToMatrix
Квантильное выравнивание (quantile normalization)
Разделение матриц
Слияние матриц
Выравнивание таблиц
Замена значений ячеек значениями из указанного файла
Замена значений ячеек заданным значением
Замена значений ячеек матрицы
Замена значений ячеек матрицы по регулярному выражению
Тест Мантеля (Mantel test)
Ранговый тест Мантеля (Rank Mantel test)
Корреляция
Преобразование Фишера
Угловое преобразование Фишера
Транспонирование
Сортировка по строке
Сортировка по столбцу
Перемножение матриц
Выборка строк таблицы
Выборка столбцов таблицы
Вставить диагональ в матрицу
Преобразовать таблицу в вектор
Преобразовать вектор в таблицу
Преобразовать вектор в таблицу по его размеру
Преобразовать вектор в диагональную матрицу

Модули (продолжение)

Записать заголовок
Переместить строку вверх
Поменять местами строки по ключу
Поменять местами столбцы по ключу
Echo
Удаление строк с нечисловыми значениями
Удаление строк, которые не содержат указанное значение в указанном столбце
Операции над элементами матрицы
Поэлементные операции с матрицами
Вычисление модуля каждого элемента
Замена ключей-строк в таблице
Замена ключей-столбцов в таблице
Замена ключей строк/столбцов числовыми значениями
Дописать таблицу справа
Дописать таблицу вниз
Дописать файл
Вычислить базовые статистики
Вычислить функцию распределения
Развертка матрицы в двоичные признаки
Разделить таблицу по значениям в заданном столбце
Разделить таблицу по значениям в заданной строке
Разделить таблицу по подстроке в ключах строк
Разделить таблицу по подстроке в ключах столбцов
Евклидова метрика
Метрика Минковского
Коэффициент Жаккара
Коэффициент Жаккара-Наумова
Расстояние Хэмминга
Манхэттенское расстояние
p-distance
Расстояние Джукса-Кантора
Расстояние Кимуры
extract
insert
Создать пустой файл
Заполнить матрицу случайными значениями
Скопировать файл
Удалить файл
Преобразовать CSV в TXT
Подсчет количества повторов ключей строк
Сдвиг матрицы

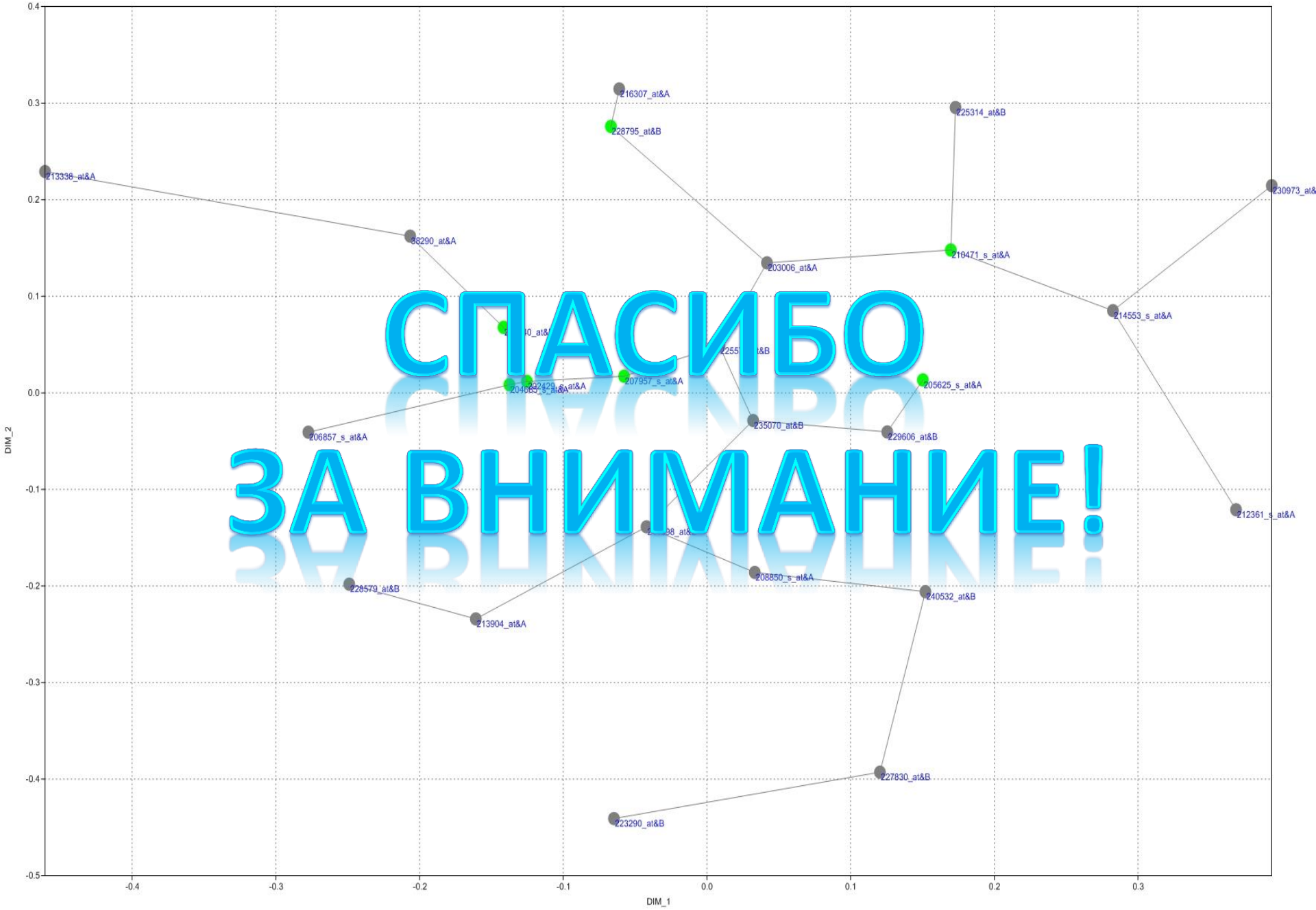
Сеть коэкспрессии генов



Цель биостатистики

Адекватная математическая обработка
статистических данных
для решения **биологических** задач

О роли пивоваров в науке


СПАСИБО
ЗА ВНИМАНИЕ!
216307_at&A
225314_at&B
228795_at&B
213336_at&A
230973_at&
38290_at&A
210471_s_at&A
203006_at&A
214553_s_at&A
207957_s_at&A
205625_s_at&A
206857_s_at&A
235070_at&B
229606_at&B
212361_s_at&A
208850_s_at&A
228579_at&B
240532_at&B
213904_at&A
227830_at&B
223290_at&B
DIM_1
DIM_2

Три регрессионных метода

Множественная линейная регрессия

Регрессия на главные компоненты

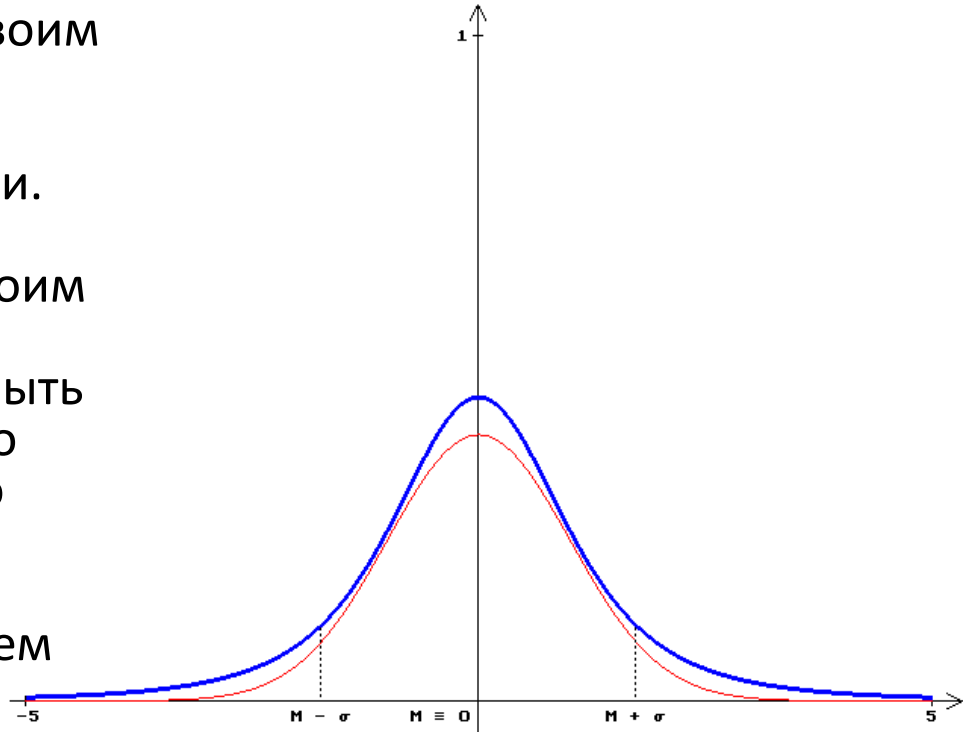
PLS - регрессия

1. Координатное представление объектов в многомерном пространстве

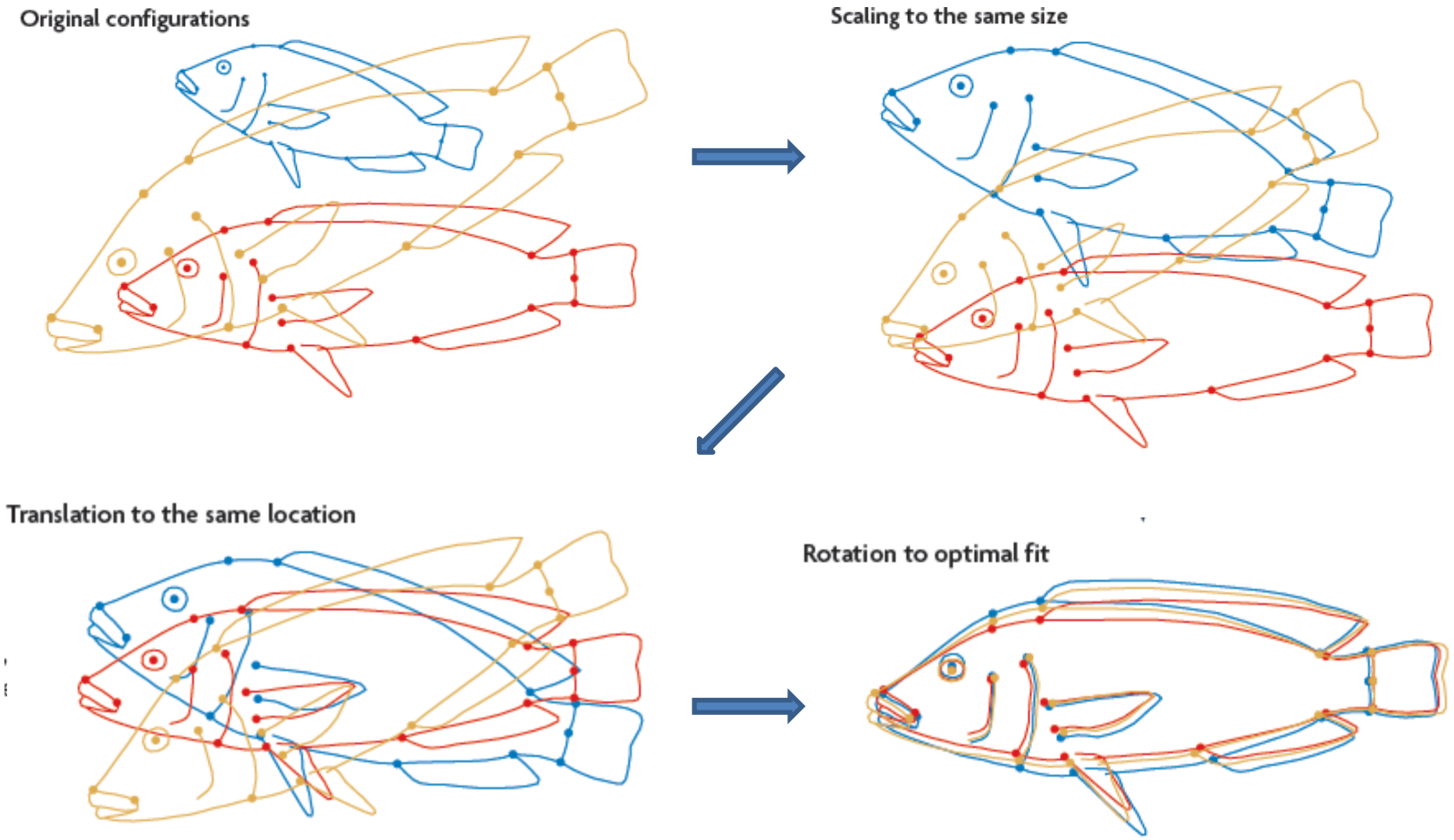
2. Матрица сходства-различия между объектами

- Интерес Госсета к выращиванию ячменя привёл его к мысли, что опыт надо планировать с той целью, чтобы не просто повысить среднюю урожайность, но чтобы вывести такие сорта ячменя, чья урожайность была бы устойчива к колебаниям состава почвы или климата. Этот принцип встречается только позднее у Фишера и затем в 50-х в работе Гэнъити Тагути.

- Чтобы предотвратить дальнейшее раскрытие конфиденциальной информации, Гиннесс запретил своим работникам публикацию любых материалов, независимо от содержавшейся в них информации. Это означало, что Госсет не мог опубликовать свои работы под своим именем. Поэтому он избрал себе псевдоним *Стьюдент*, чтобы скрыть себя от работодателя. Поэтому его самое важное открытие получило называние Распределение Стьюдента, иначе бы оно могло называться теперь распределением Госсета.



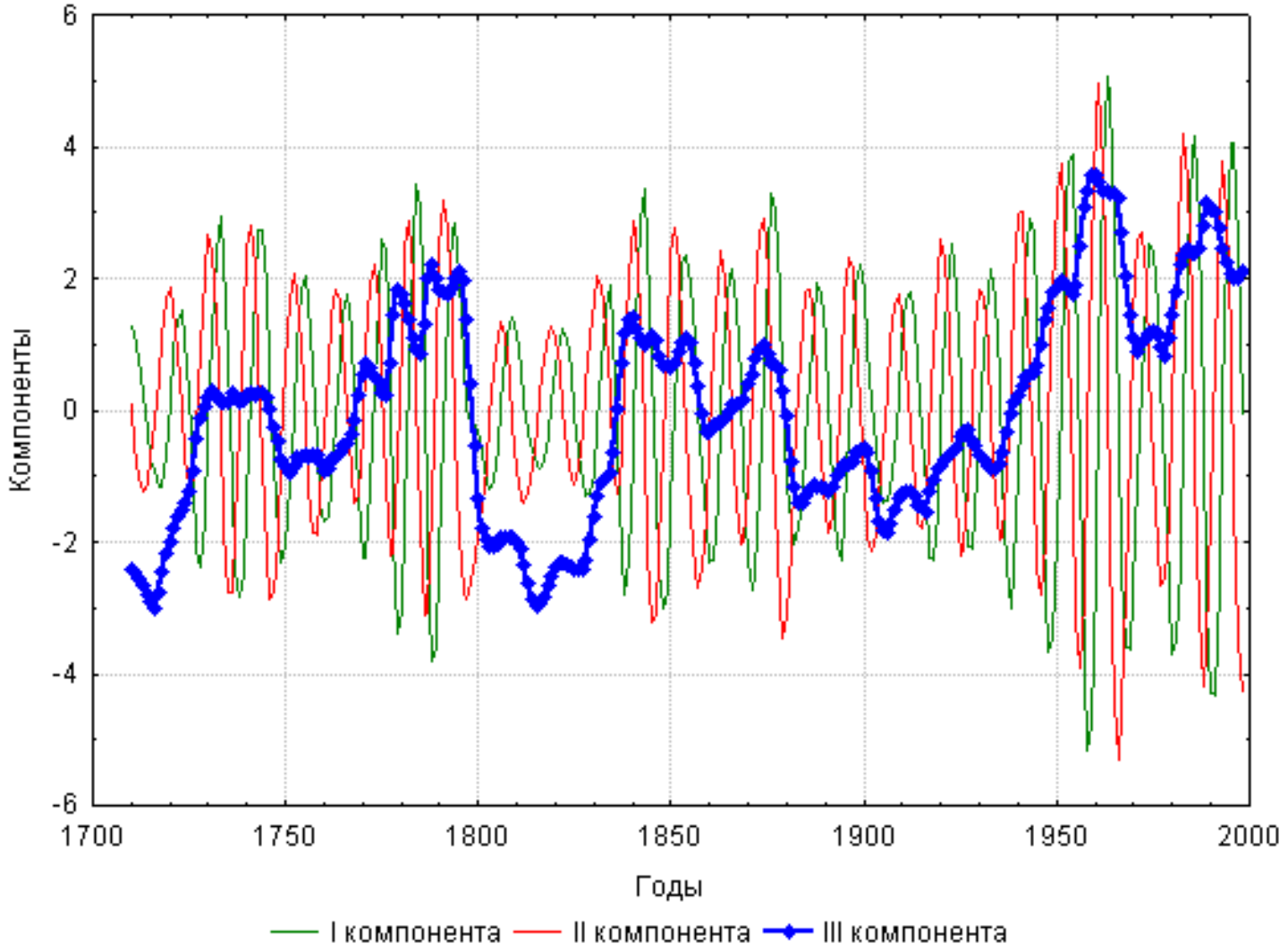
2B-PLS анализ



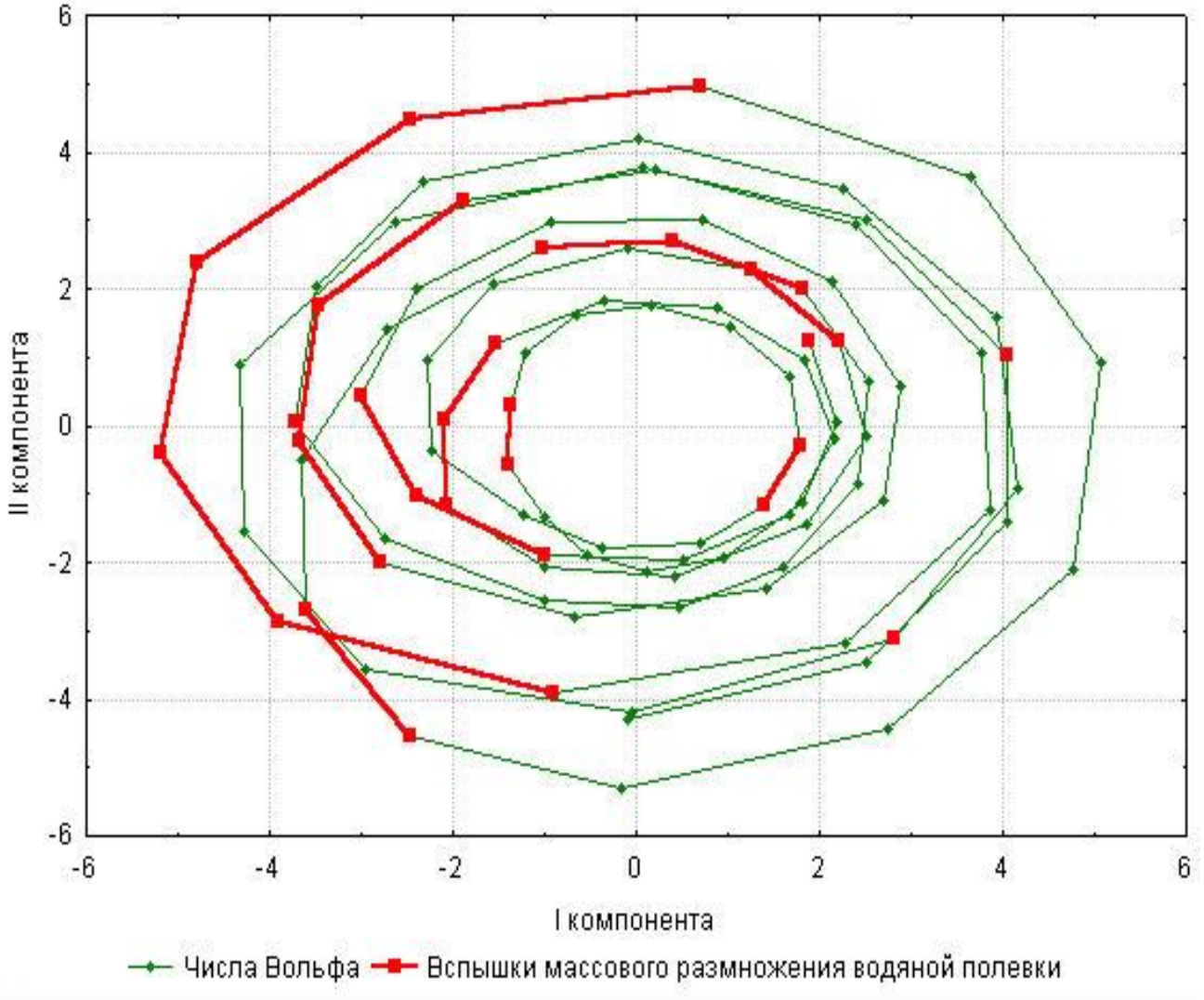
Цель – добиться максимального соответствия расположения меток.
Однако метод работает и в общем случае.

V. Viscosi, A. Loy, P. Fortini. (2010). Geometric morphometric analysis as a tool to explore covariation between shape and other quantitative leaf traits in European white oaks. In: Nimis P. L., Vignes Lebbe R. (eds.). Tools for Identifying Biodiversity: Progress and Problems – pp. 257-261.

Первые три главные компоненты ряда Вольфа



Траектория ряда Вольфа в фазовом пространстве I-II главных компонент и годы вспышек массового размножения водяной полевки



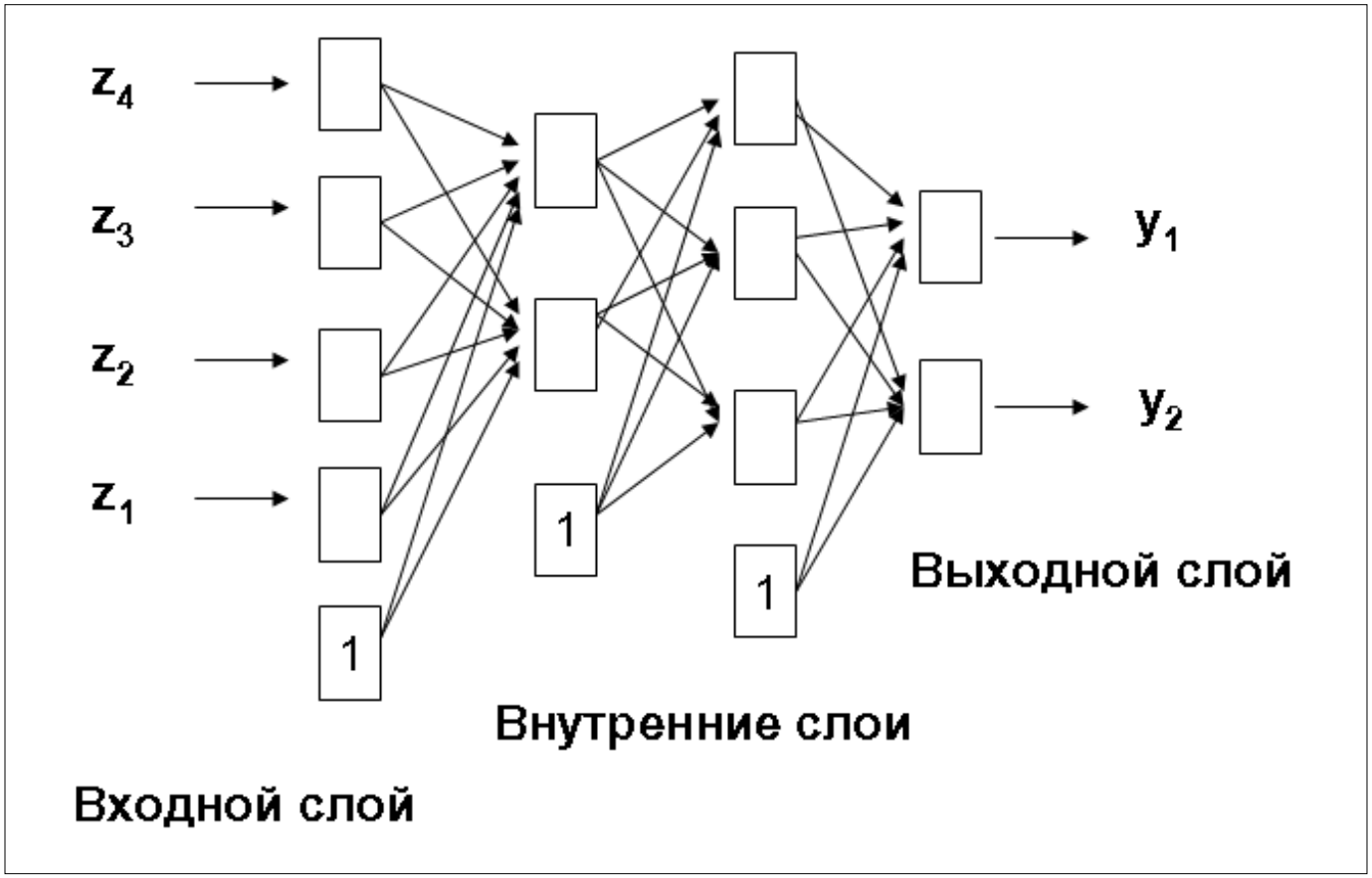
Многомерное шкалирование

Предполагая, что эксперт может оценить различия между парами объектов настолько, что можно их упорядочить, можно поставить задачу определения координат объектов в многомерном пространстве с заданной метрикой (удобнее всего, евклидовой) таким образом, чтобы ранги различий как можно ближе соответствовали рангам дистанций между этими же парами в многомерном пространстве. Эти соображения легли в основу дистанционной модели М.Ричардсона (Richardson, 1938) – первого варианта неметрического многомерного шкалирования. Однако, из-за отсутствия вычислительных возможностей в то время этот метод не мог быть реализован. Поэтому В.Торгерсон предложил рассматривать различия между парами объектов как прямые аналоги расстояний в многомерном пространстве и разработал метод, позволяющий приписывать объектам координаты с сохранением расстояний – метрическая модель Торгерсона (Torgerson, 1952; Торгерсон, 1972). Эту модель уже можно было реализовать на компьютерах, что и было сделано. Но ее условия применимости оказались слишком жесткими, многие меры близости, применяемые психологами, явно не соответствовали аксиомам метрического расстояния, поэтому Р.Шепард и Дж.Крускал вернулись к первоначальным предположениям дистанционной модели М.Ричардсона.

Нейронные сети

В качестве нелинейного обобщения множественной регрессии можно рассматривать некоторые варианты нейронных сетей. Нейронная сеть является крайне упрощенной вычислительной моделью человеческого мозга и состоит из нейронов, соединенных друг с другом. Одна часть нейронов воспринимает входную информацию, другая работает на выдачу результатов, остальные скрыты от внешнего наблюдателя.

Нейронные сети



СПАСИБО ЗА ВНИМАНИЕ!

Среда 12.12. 16:00
Конференц-зал ИЦИГ

Публичная лекция
В.М.Ефимова

Зачем
БИОлогу
БИОстатистика