Многомерные методы анализа данных и моделирования медицинских систем

Долецкий А.Н.

ПЛАН

1. Многомерные методы анализа данных
2. Задачи исследования сложных систем.
3. Дискриминантный анализ.
4. Анализ временных рядов. Кластерный анализ.

Определение

- Методы многомерного анализа (англ. multivariate или multivariable analysis) разработаны для оценки влияния более чем одного фактора на результат, а также об эффекте взаимодействия этих переменных между собой.
- В отечественной литературе многомерный анализ часто называют многофакторным анализом.

Пример

- Пациенты травматологического отделения, оперированные в 9 часов утра, имеют более высокий показатель смертности, чем пациенты, оперированные в 9 часов вечера.
- Анализируя смертность после операции одномерным (параметрическим или непараметрическим) статистическим методом, исследователь, допустим, действительно может обнаружить достоверную разницу.

Пример

- Время суток не является единственным определяющим фактором.
- В анализ должны быть включены тяжесть травмы, возраст больного, плановые или срочные показания к операции, и т. п.

После этого вклад времени операции окажется очень малым по сравнению с истинными причинами.

Терминология

- Факторы (причины), влияющие на исход - факторы риска (risk factors), независимые (independent) или объясняющие переменные (explanatory variable)
- Исход (outcome) – зависимая (dependent) или переменная отклика (response variable) или эффект.

NB

- Факторы риска — не воздействия в процессе эксперимента!
- Экспериментальная проверка совместного влияния многих факторов в клинической практике чаще всего просто невозможна или недопустима по этическим соображениям.
- *Повышает ли курение вероятность ИБС в двух случайных выборках людей, одни из которых курят, другие – нет?*

- На связь между курением и ИБС могут влиять мешающие факторы (confounders).

?

- многомерный анализ является не единственным статистическим методом учета влияния или исключения конфаундеров.
- Оценить влияние фактора риска на исход можно также с помощью **условного анализа -** изучаемая группа разбивается на подгруппы (страты), в которых потенциально мешающая переменная «фиксируется».
- При этом модель применяется отдельно на каждой группе данных.
- Например, мужской и женский пол: только курящие или некурящие мужчины и только курящие или некурящие женщины.
- **Условный анализ эффективен когда изучается относительно небольшое число факторов** (два-три).

Регрессия

В зависимости от задачи и типов данных чаще всего используется

- множественная линейная регрессия (multiple linear regression),
- множественная логистическая регрессия (multiple logistic regression) и
- модель пропорциональных интенсивностей Кокса (Cox proportional hazards model).

Множественная линейная регрессия

- Независимые переменные - **непрерывные численные интервальные или относительные**.
- Значения независимых переменных и зависимой переменной (исхода) изменяются линейно.
- Величина коэффициента при независимой переменной и его знак в конечной модели показывают степень и характер взаимосвязи между этой переменной и исходом.

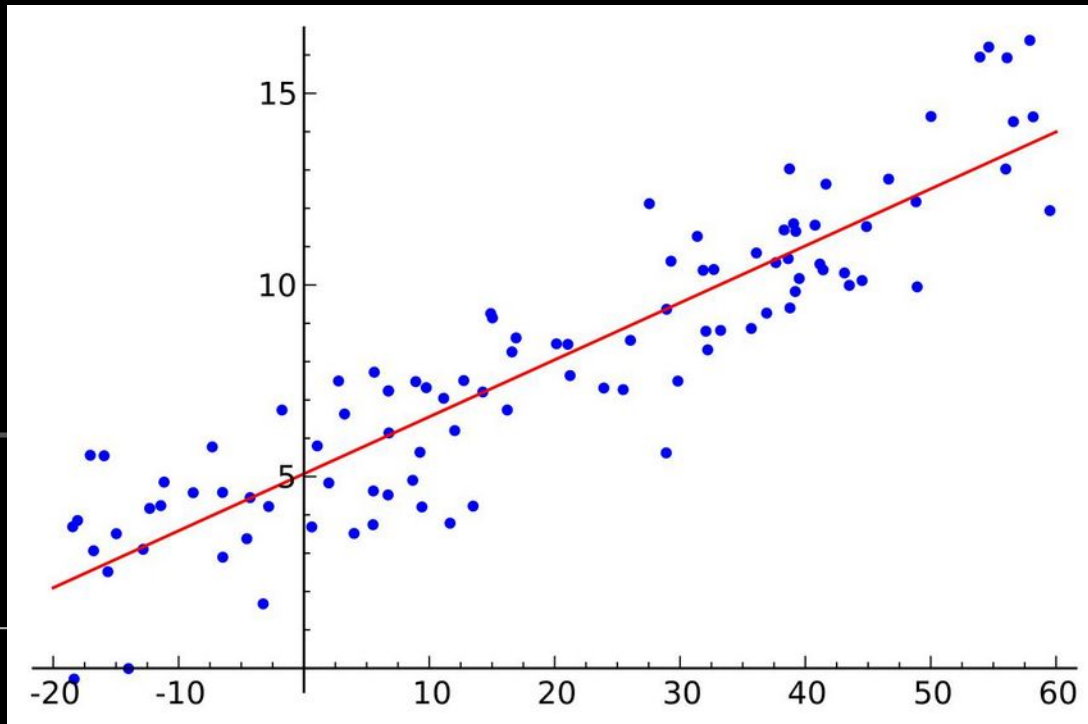
Множественная линейная регрессия

- модель оценки костной плотности у женщин в менопаузе, в которую входят возраст и индекс массы тела.

Предназначение создаваемой модели

- **Объясняющая** модель.
- **Прогностическая** модель.

Могут быть включены коррелирующие независимые переменные, и модель будет хорошо работать. Но оцененные коэффициенты модели для таких переменных не имеют самостоятельного значения и не несут информации о степени влияния на эффект.



Анализ качества модели

- Для оценки эффективности множественной линейной регрессии используется коэффициент детерминации r2.
- Отражает степень рассеяния результата, возникающего благодаря вкладу многих переменных. Значение r2 варьирует в пределах от 0 до 1 и чем ближе оно к 1, тем лучше модель описывает результат.

Логистическая регрессия

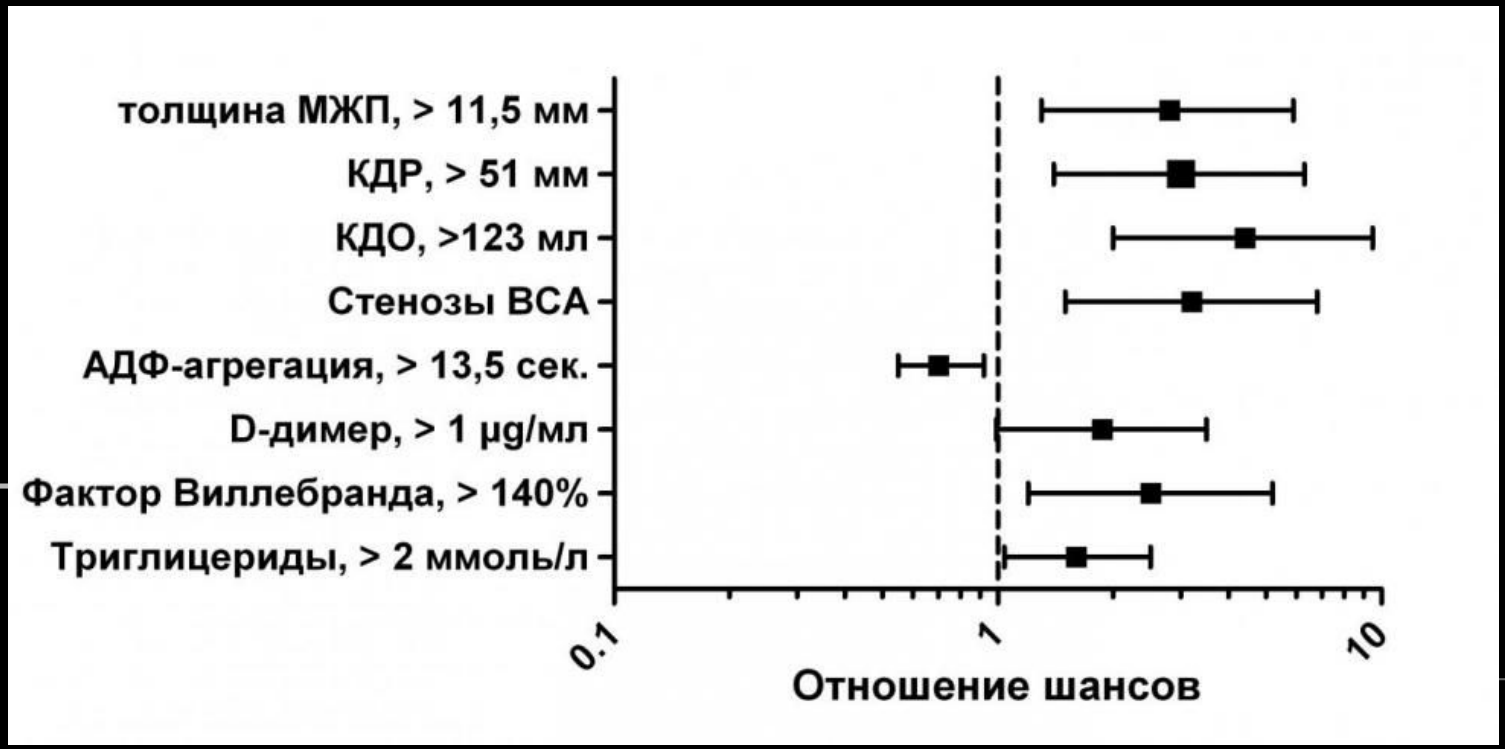
Значение переменной результата является **бинарным**:

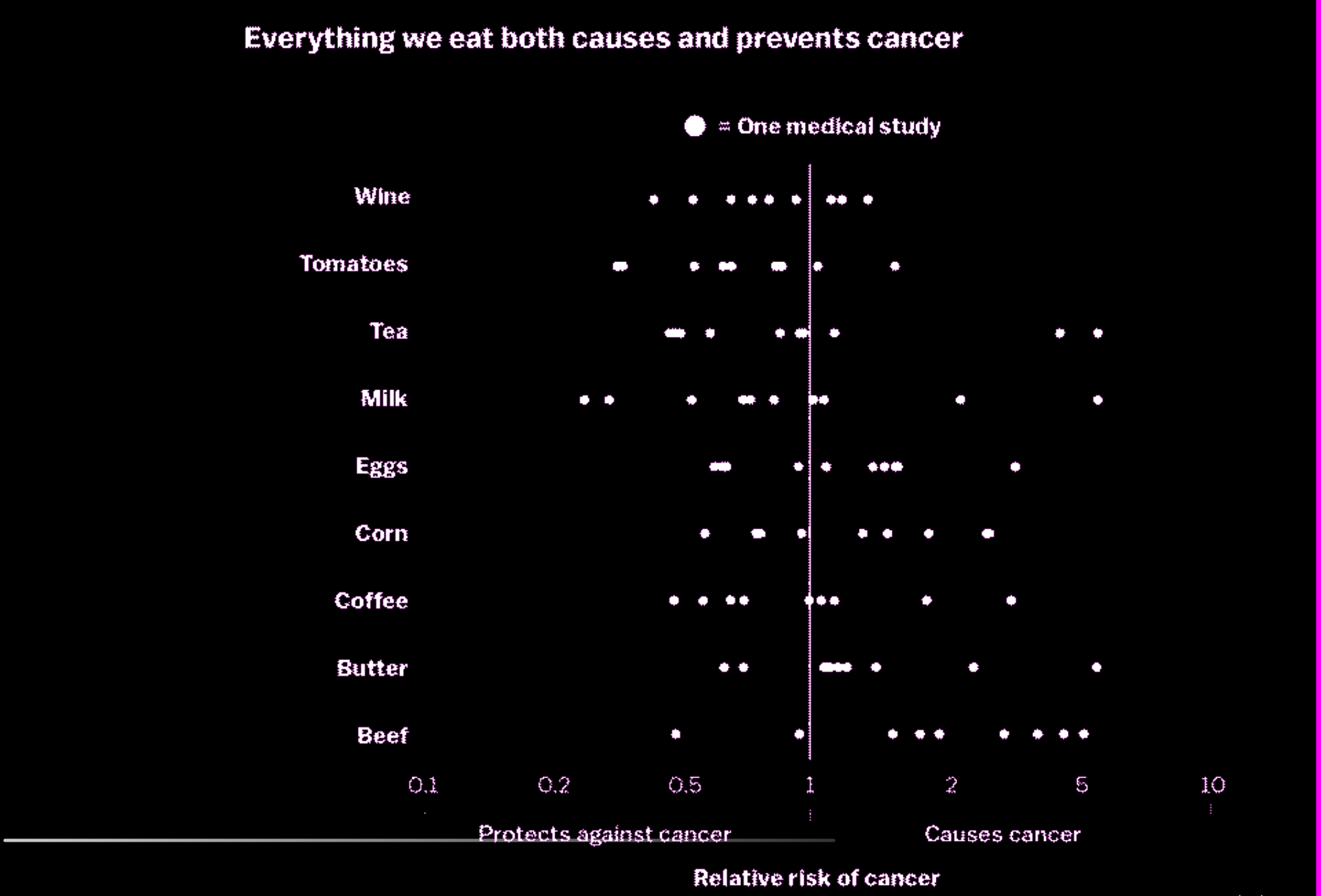
- выживаемость (да/нет),
- развитие заболевания (да/нет),
- положительный результат диагностического теста (да/нет).

Относительный риск (relative risk – RR), **отношение шансов** (odds ratio - OR) и границы их доверительных интервалов (confidential intervals - CI), а также степень достоверности (величину p) отличия этих величин от 1 (значение при нулевой гипотезе).

Логистическая регрессия

OR близкое к 1 свидетельствует о слабой взаимосвязи. Широкий CI — низкая надежность оценки и необходимость перепроверки данных. **Если границы CI включают значение 1, связь независимой переменной с исходом не может быть признана достоверной.**





Ioannidis, Why most published research findings are false, 2005.

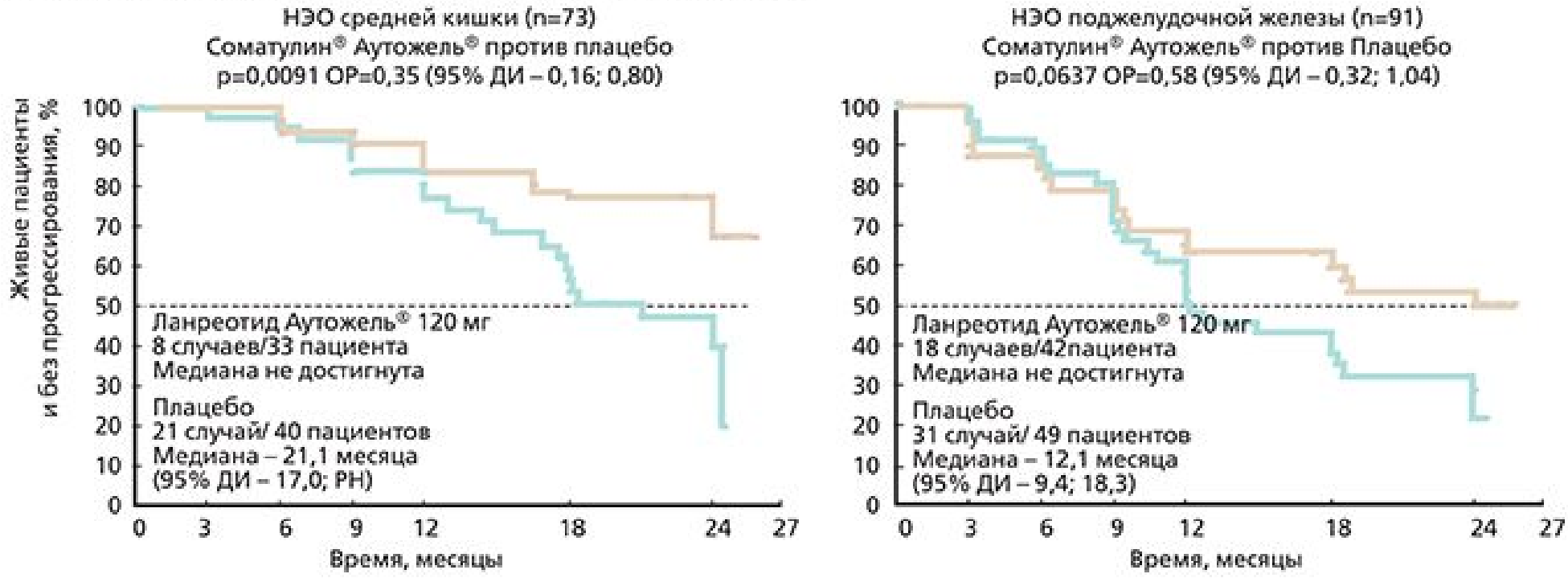
Модель пропорциональных рисков Кокса

Зависимая переменная представляет время наступления определенного события.

Оценивает шансы более раннего наступления события у членов изучаемой группы по сравнению с контрольной группой с помощью показателя отношения рисков (HR).

HR=2 — пациент, получающий препарат и у которого выздоровление еще не наступило, имеет в 2 раза больший шанс выздороветь к следующему моменту времени, чем тот, кто получал плацебо.

Модель пропорциональных рисков Кокса



Значение p получено на основании стратифицированного log-рангового критерия;
ОР (отношение рисков) – получено на основании модели пропорциональных рисков Кокса.

Примечание. РН – рассчитать невозможно.

Caplin M., et al. North American Neuroendocrine Tumor Symposium, Charleston, South Carolina, USA, 2013.

ФАКТОРНЫЙ АНАЛИЗ

Существует латентная переменная, с помощью которой можно объяснить наблюдаемое сходство полученных оценок. Такую латентную переменную называют фактором, который влияет на показатели других переменных.

цели факторного анализа:

1. определение взаимосвязей между переменными, их классификация, т. е. «объективная R-классификация»;
2. сокращение числа переменных.

	Фактор					
	1	**2**	**3**	**4**	**5**	**6**
Результаты фрактального (энтропийного) анализа						
индекс Херста	-0,03	0,04	-0,14	-0,28	0,15	**0,78**
индекс Дельта	0,04	-0,04	0,02	0,22	-0,11	**0,71**
Средняя амплитуда ритмов в частотных диапазонах						
Дельта	0,63	0,28	-0,60	-0,05	-0,01	0,03
Тета	**0,87**	0,27	0,01	0,01	0,01	-0,09
Альфа	**0,76**	0,22	0,31	0,00	0,17	0,10
Бета-1	**0,91**	-0,14	0,06	0,18	0,01	-0,04
Бета-2	**0,82**	-0,38	-0,12	0,10	0,02	0,15
Результаты автокорреляционного анализа						
Средняя частота	0,00	**-0,93**	0,05	0,07	0,03	-0,02
Интервал корреляции	-0,17	0,03	**-0,90**	-0,07	0,10	0,06
Коэффициент корреляции	0,14	0,43	**0,74**	-0,10	0,00	0,16
Результаты кросскорреляционного анализа						
Средняя частота	-0,07	**-0,95**	-0,14	-0,01	-0,08	0,04
Задержка	-0,07	0,07	0,24	0,19	**-0,71**	0,17
Коэффициент корреляции	0,08	0,13	0,20	0,29	**0,81**	0,21
Результаты когерентного анализа						
Средняя когерентность	0,01	-0,05	0,02	**0,87**	0,39	0,03
Доминирующая частота	0,12	0,00	-0,57	0,22	-0,06	0,32
Средняя частота	0,19	-0,06	-0,18	**0,81**	-0,30	-0,01

Дискриминантный анализ

- Дискриминантный анализ проводится с целью выявления различий между исследуемыми группами.
- *Цель анализа — выяснить, можно ли составить «типичный портрет покупателя» для каждой исследуемой группы по заданным характеристикам.*

Дискриминантный анализ

- Система уравнений дискриминантного анализа:
- ***Si = ci + wi1*x1 + wi2*x2 + … + wim*xm***

Si - результат показателя классификации;

i – номер группы;

индексы 1, 2, …, m – переменных (номер исследования);

ci - константа для i-ой группы;

xj — наблюдаемое значение j-ой переменной.

Дискриминантный анализ

```
## [1] "LDA - зависимость ESS от качества сна, тревожности и депрессии"
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.11995    0.94663  14.916  < 2e-16 ***
## Сумма       -0.32398    0.04428  -7.318 5.33e-13 ***
## Сумма_3      0.12650    0.05064   2.498   0.0127 *
## Сумма_4      0.11066    0.06730   1.644   0.1005
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.75 on 961 degrees of freedom
##   (56 observations deleted due to missingness)
## Multiple R-squared:  0.1113, Adjusted R-squared:  0.1085
## F-statistic:  40.1 on 3 and 961 DF,  p-value: < 2.2e-16
```

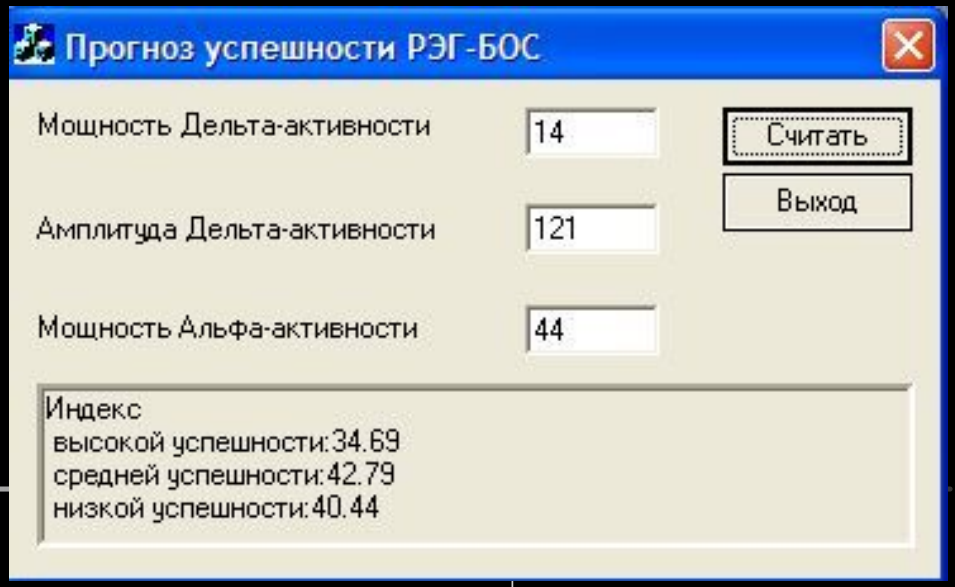
РАННЯЯ ДИАГНОСТИКА РАЗВИТИЯ ТОКСИКОЗА

- норма = 0.865СрАД - 0.126СИ + 1.2ФВ - 71.1
- гестоз = СрАД - 0.168СИ + 1.1ФВ - 77.446

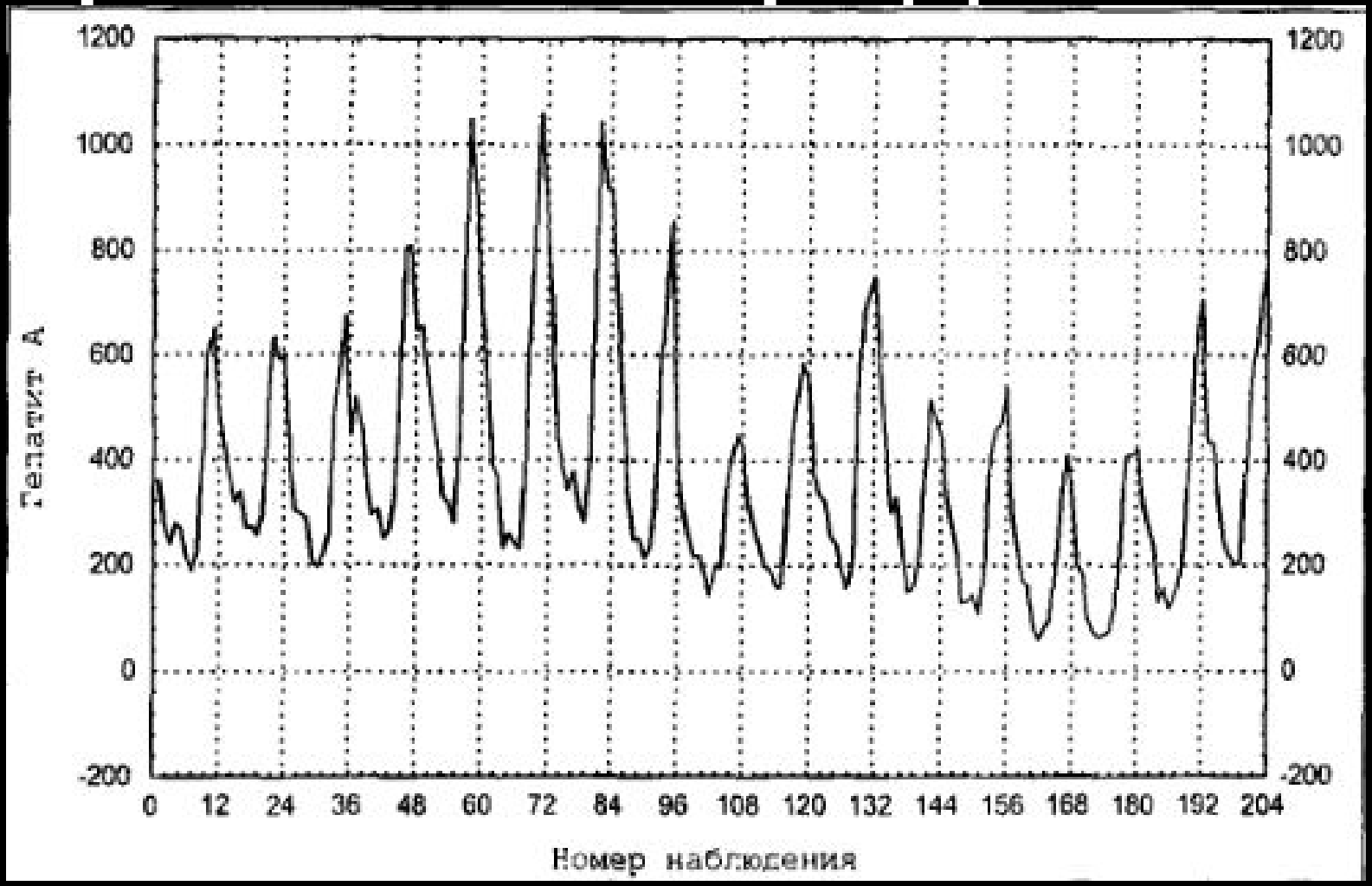
Выделение групп с различной успешностью обучения

Весовые коэффициенты параметров при вычислении показателя классификации для групп с различной успешностью обучения БОС

Исследование	параметр	группа 1	2	3
ЭЭГ	Мощность дельта-ритма	2,40	2,23	2,37
ЭЭГ	Амплитуда дельта-ритма	-0,09	0,16	-0,01
ЭЭГ	Мощность альфа-ритма	2,24	2,05	2,20
Константа		-86,9	-97,9	-88,4

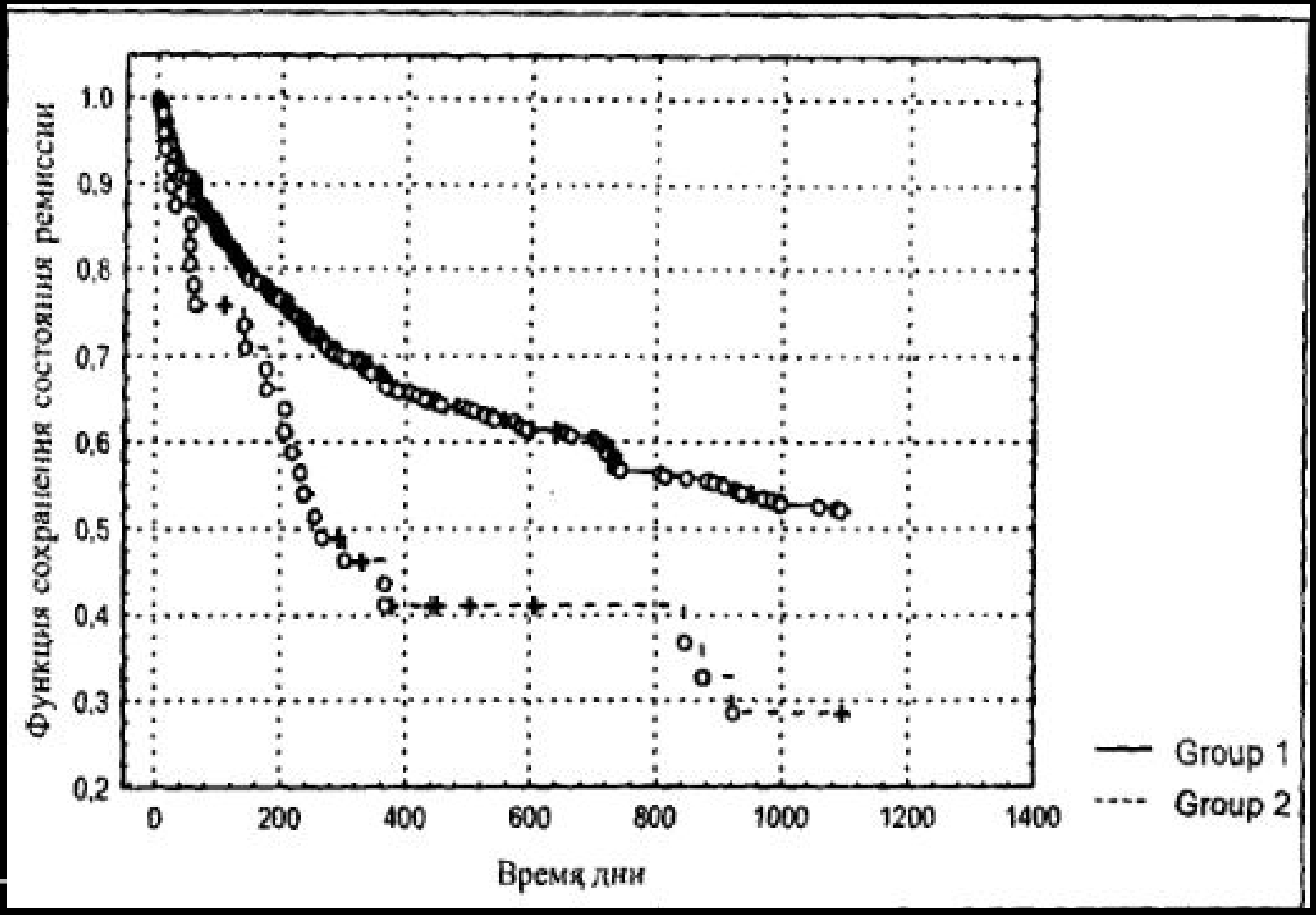


Временные ряды



Помесячное количество случаев гепатита А

Временные ряды



Сохранность ремиссии в зависимости от вида платы за лечение

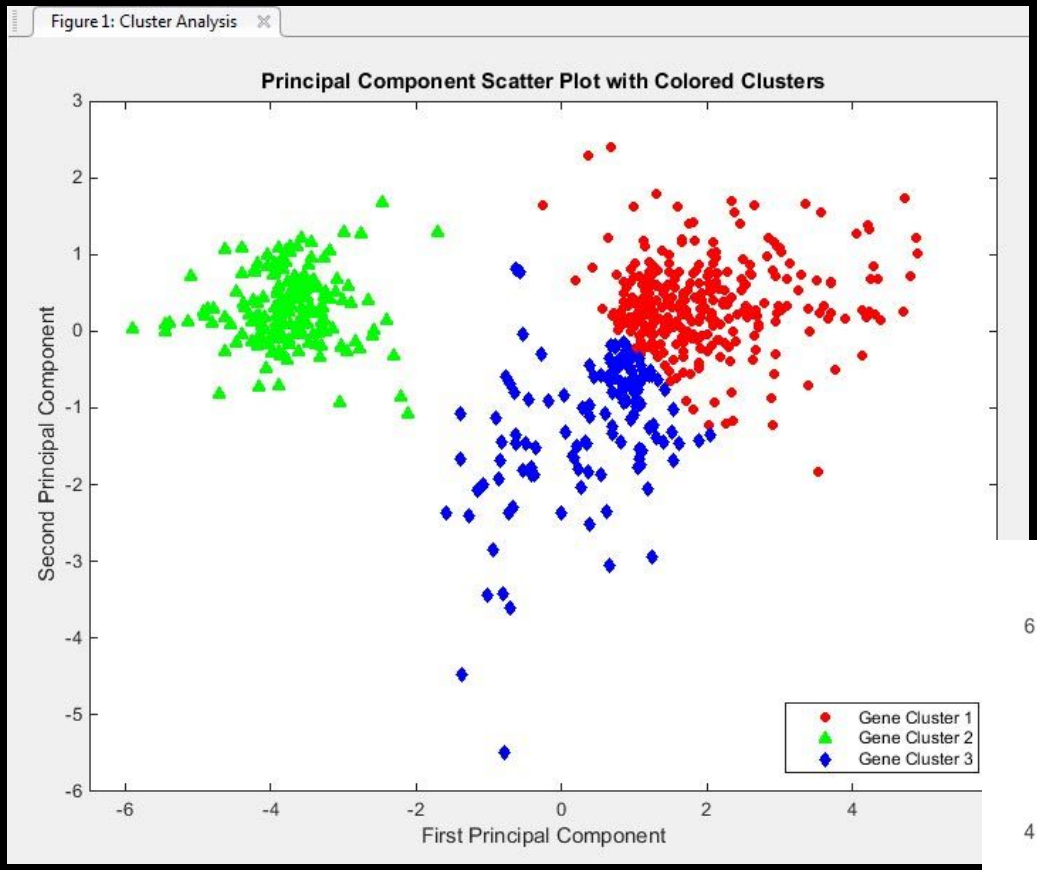
Кластерный анализ

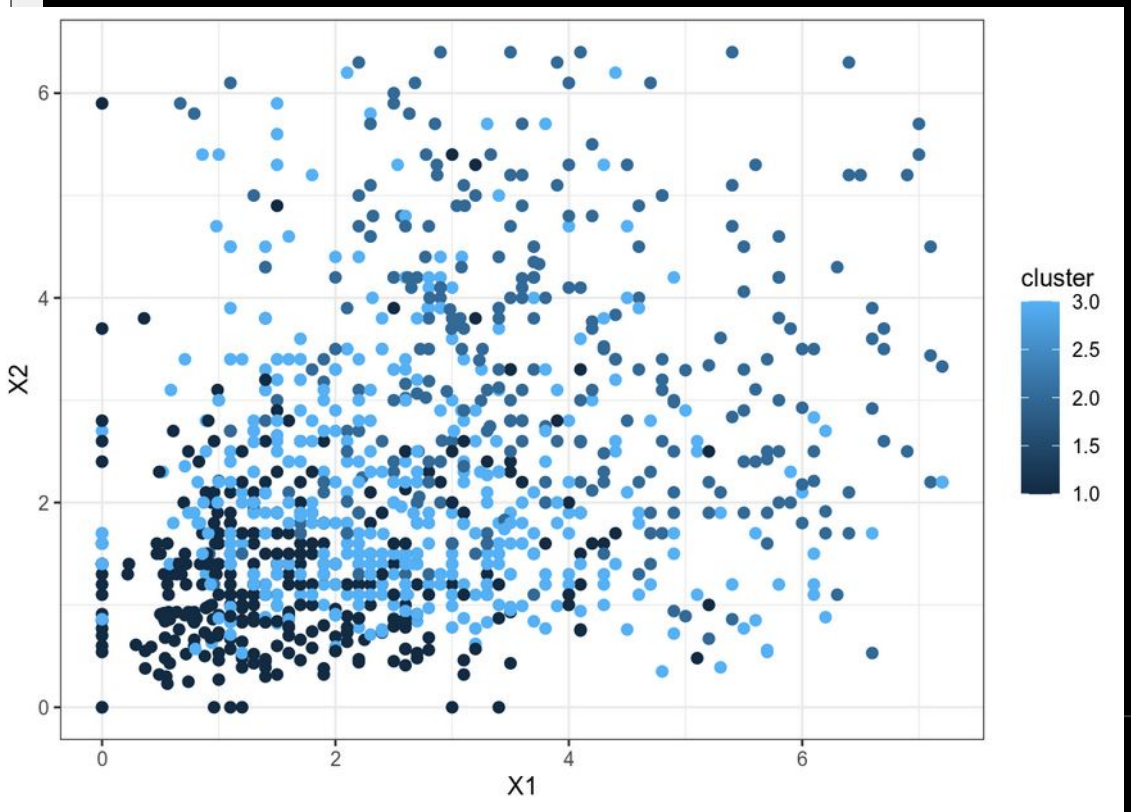
Кластерный анализ — метод классификации объектов по заданным признакам. Задача кластерного анализа состоит в формировании групп:

- однородных внутри (условие внутренней гомогенности);
- четко отличных друг от друга (условие внешней гетерогенности).

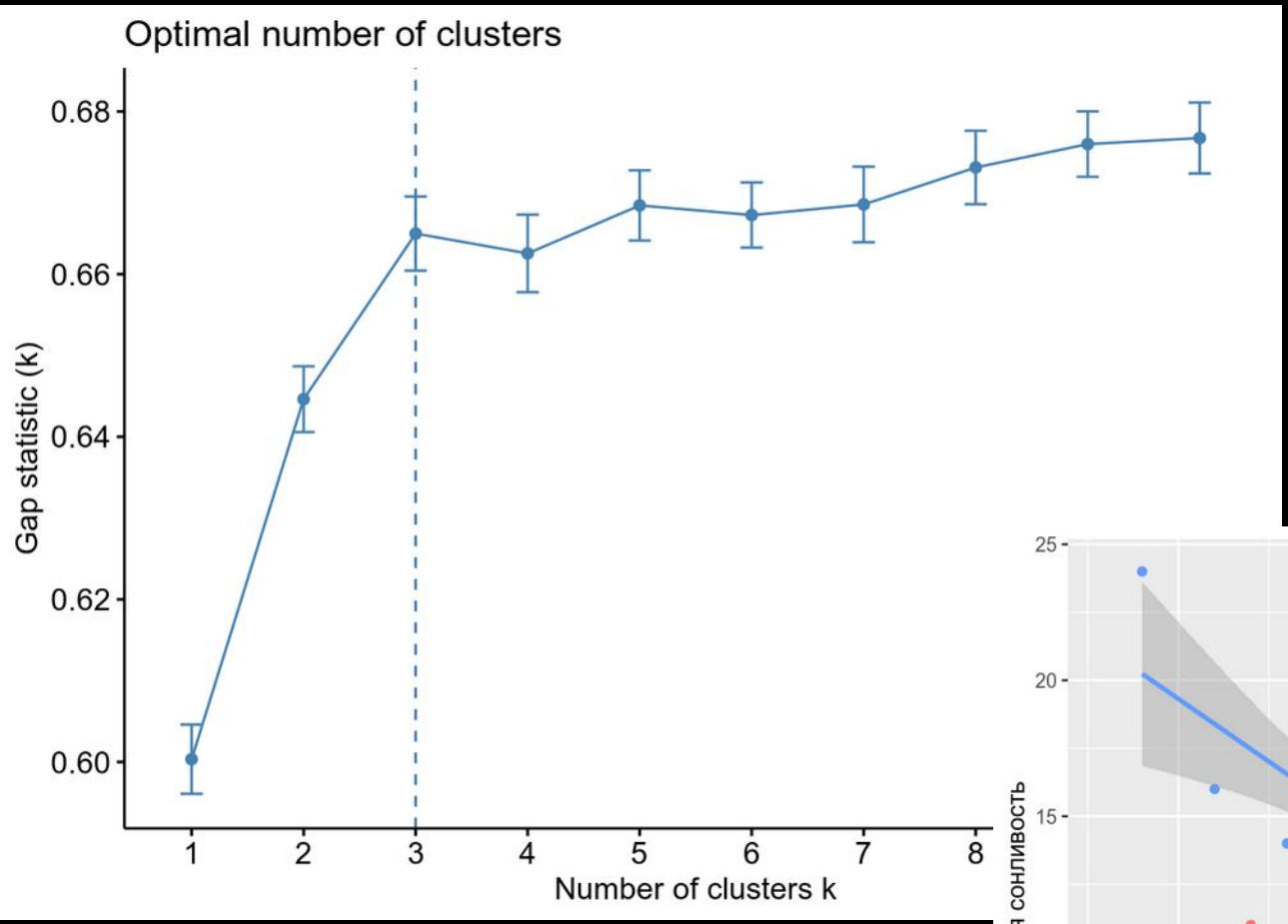
Техника кластерного анализа заключается в выявлении уровня схожести всех исследуемых элементов и последовательном объединении элементов в порядке возрастания уровня различия между ними. Число выявленных кластеров зависит от заданного уровня схожести (различия) элементов, включаемых в один кластер.

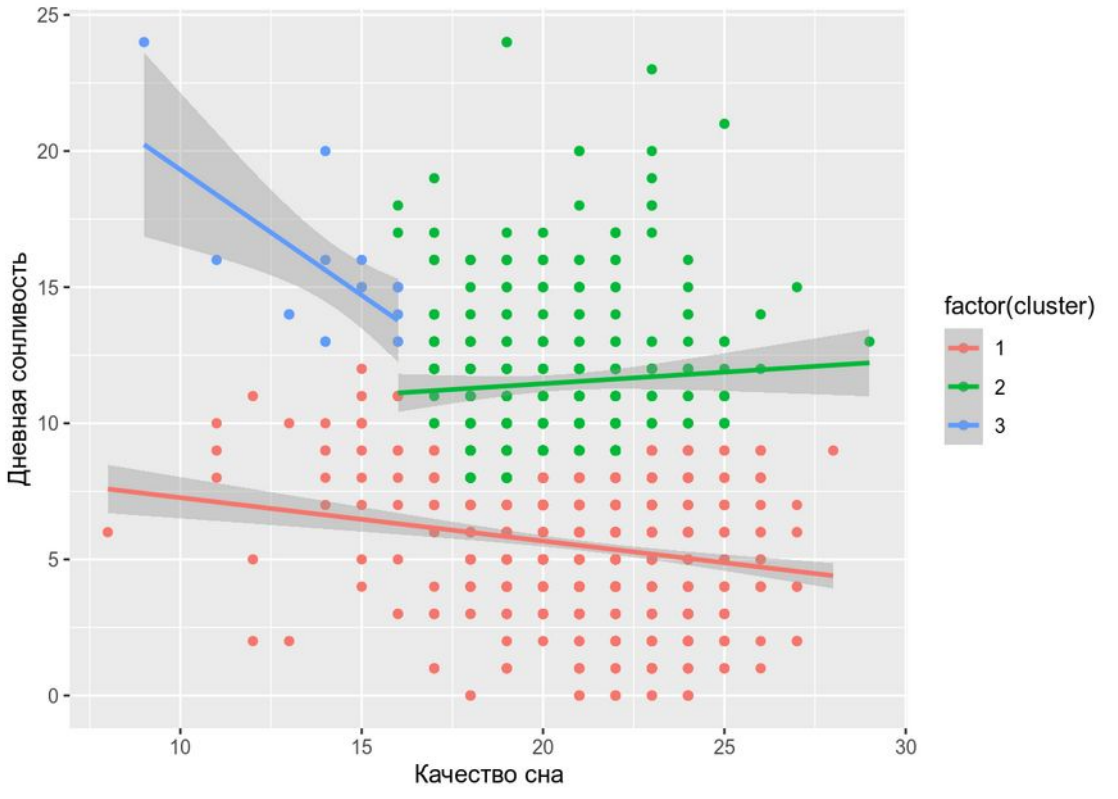
Кластерный анализ



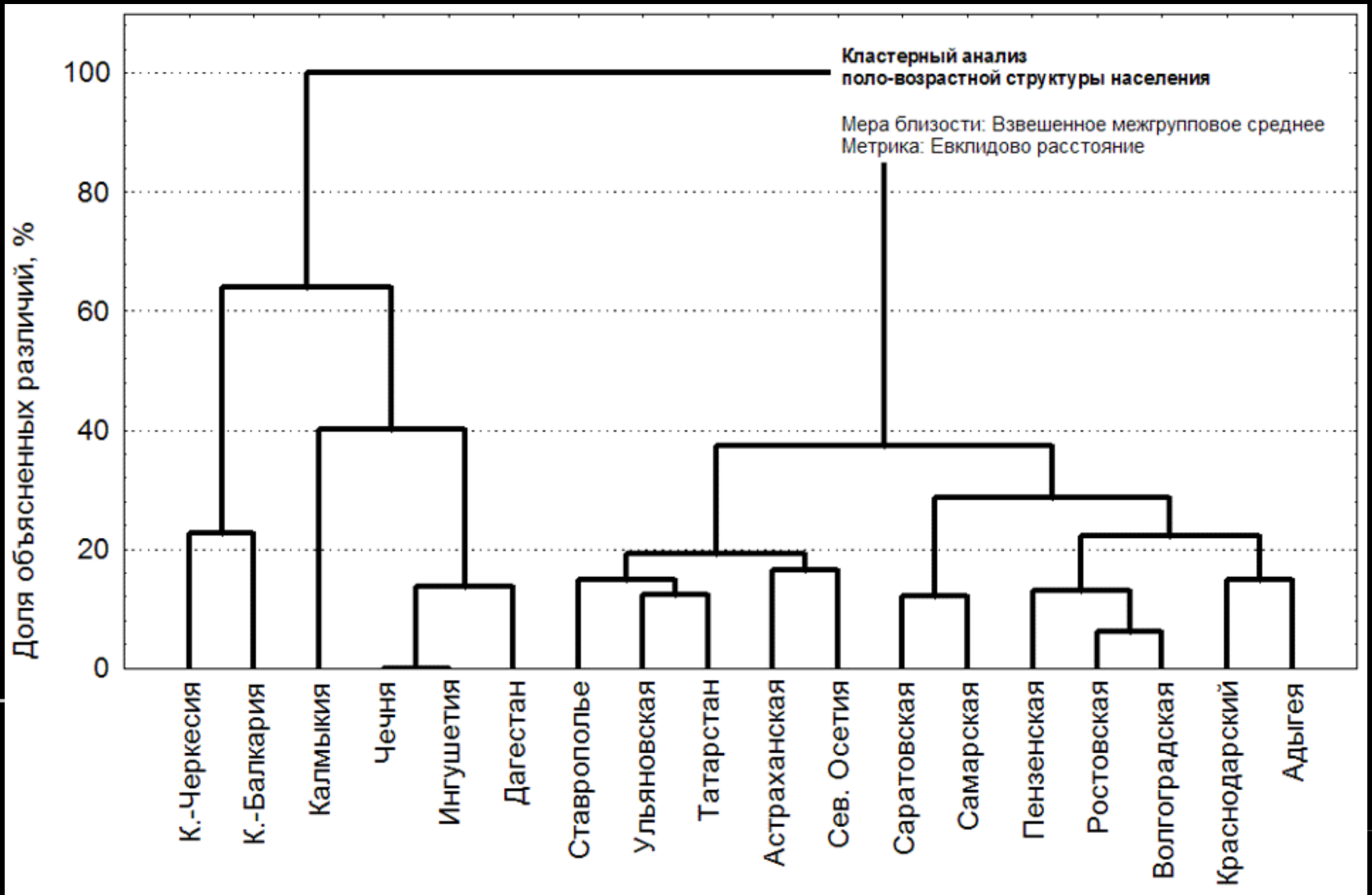


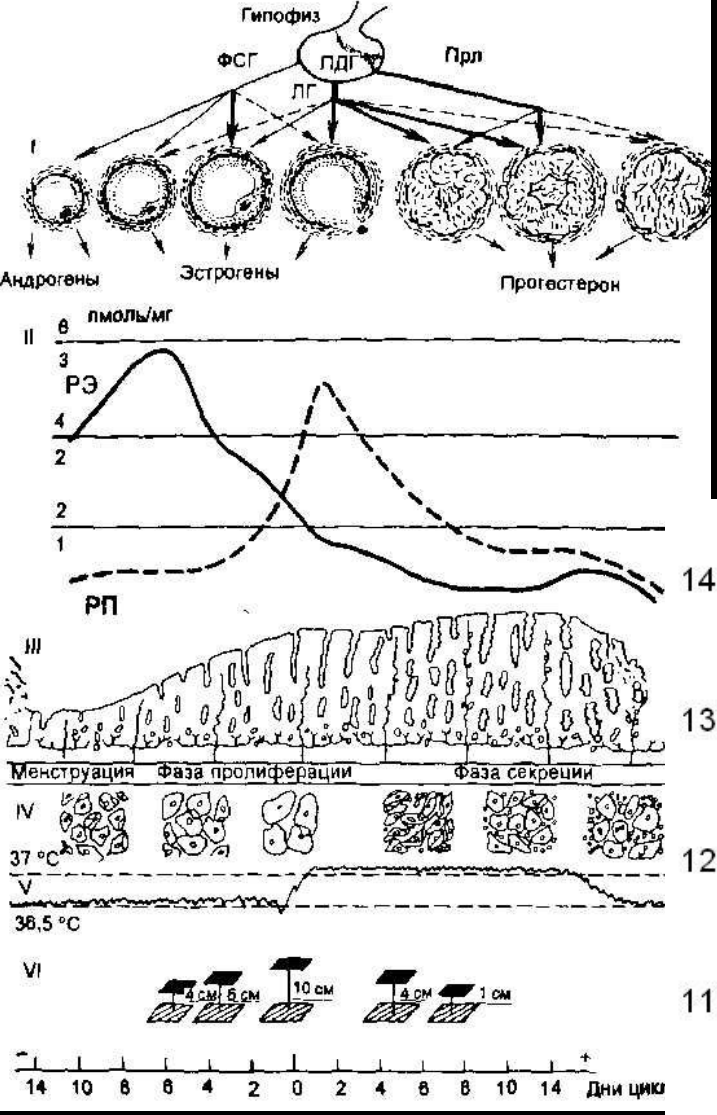
Кластерный анализ

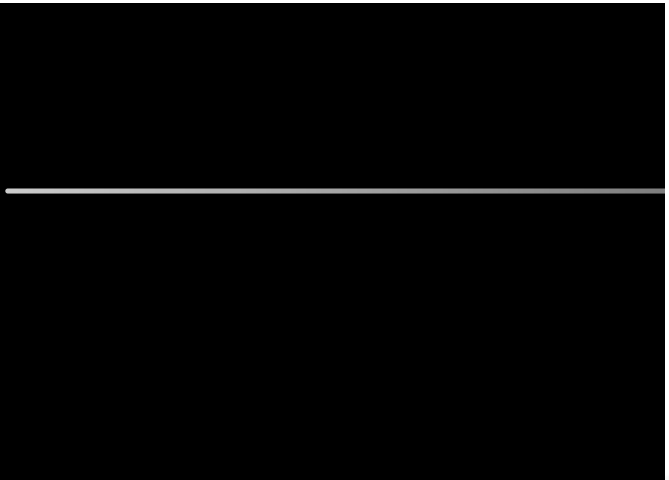


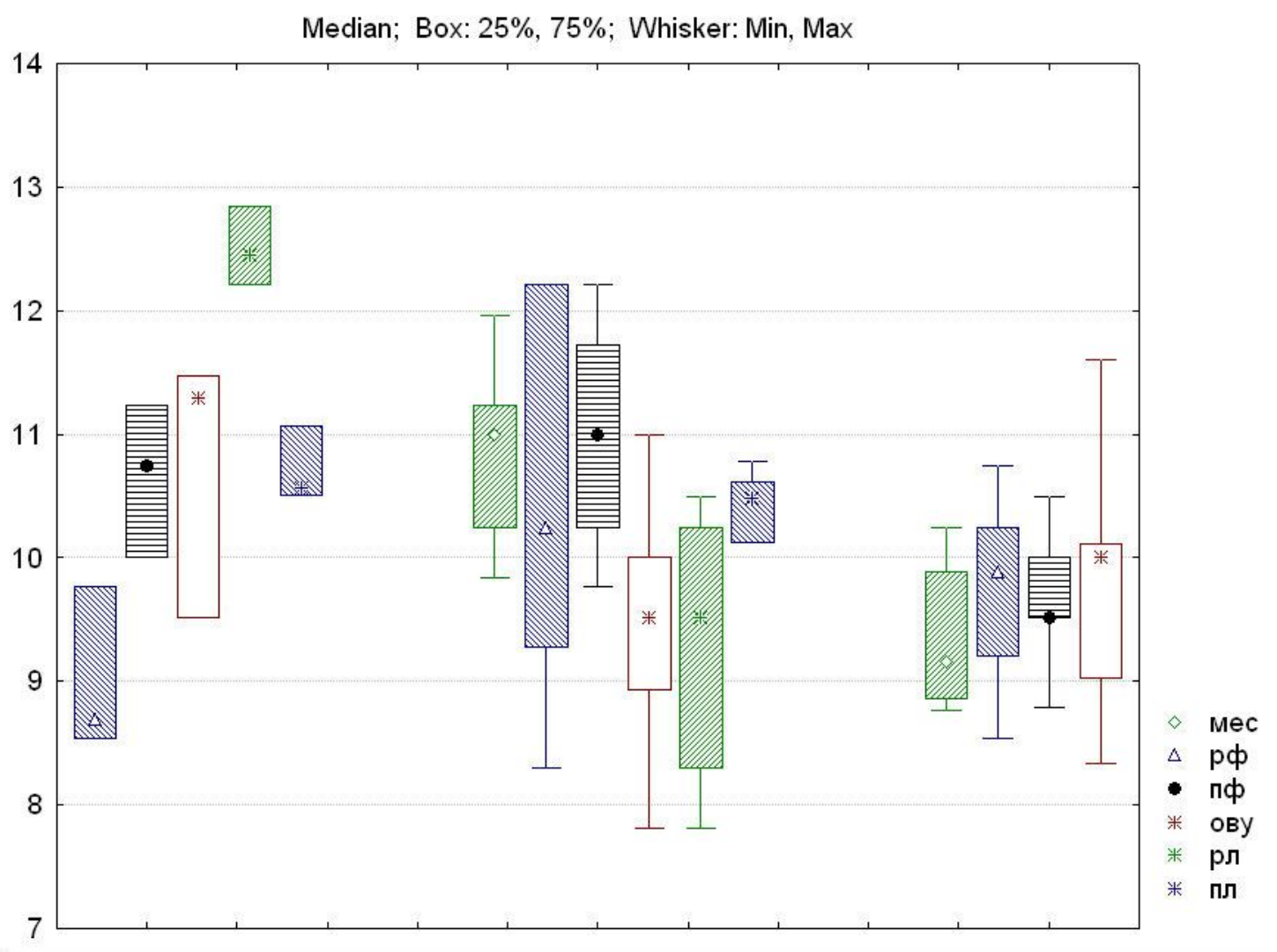


Кластерный анализ




Гипофиз
ФСГ
ЛГ
Прл
I
Андрогены
Эстрогены
Прогестерон
II
пмоль/мг
РЭ
РП
III
Менструация
Фаза пролиферации
Фаза секреции
IV
37 °C
V
36,5 °C
VI
4 см
6 см
10 см
4 см
1 см
14 10 8 6 4 2 0 2 4 6 8 10 14
Дни цик




Median; Box: 25%, 75%; Whisker: Min, Max
14
13
12
11
10
9
8
7
мес
рф
пф
ову
рл
пл

- В том случае, когда нет возможности получить истинные повторности наблюдений, разработаны методы, которые формируют так называемые "псевдовыборки".
- Методы "численного ресамплинга" {resampling} или, как их иногда называют в русскоязычной литературе, "методы по взятию повторных выборок" объединяют три разных подхода, отличающихся по алгоритму, но близких по сути: метод "складного ножа" {jackknife}, бутстреп {bootstrap} и метод перестановок {permutation}.

ВОЕННО-МЕДИЦИНСКАЯ АКАДЕМИЯ

В.И.Юнкеров, С.Г.Григорьев

МАТЕМАТИКО-СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Санкт-Петербург
2002