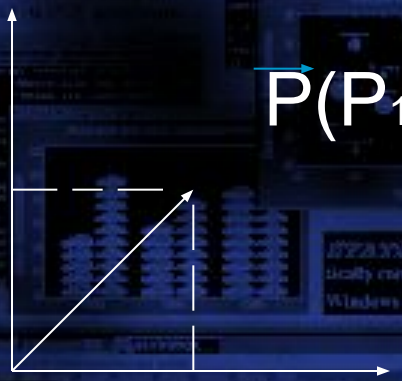
Параметрические и непараметрические методы статистики

ВВЕДЕНИЕ

Вектор состояния $\vec{P}$ (P_1,P_2,P_3…P_n) – набор функциональных параметров организма, который позволяет описать его состояние в любой момент времени.

- Пространство состояний – координатное пространство, по осям которого отложены функциональные параметры.



Среднее и доверительный интервал. Вероятно, большинство из вас использовало такую важную описательную статистику, как среднее. Среднее - очень информативная мера "центрального положения" наблюдаемой переменной, особенно если сообщается ее доверительный интервал.

Доверительный интервал для среднего представляет интервал значений, где с данным уровнем доверия находится "истинное" (неизвестное) среднее популяции.

Определение вектора состояния в норме

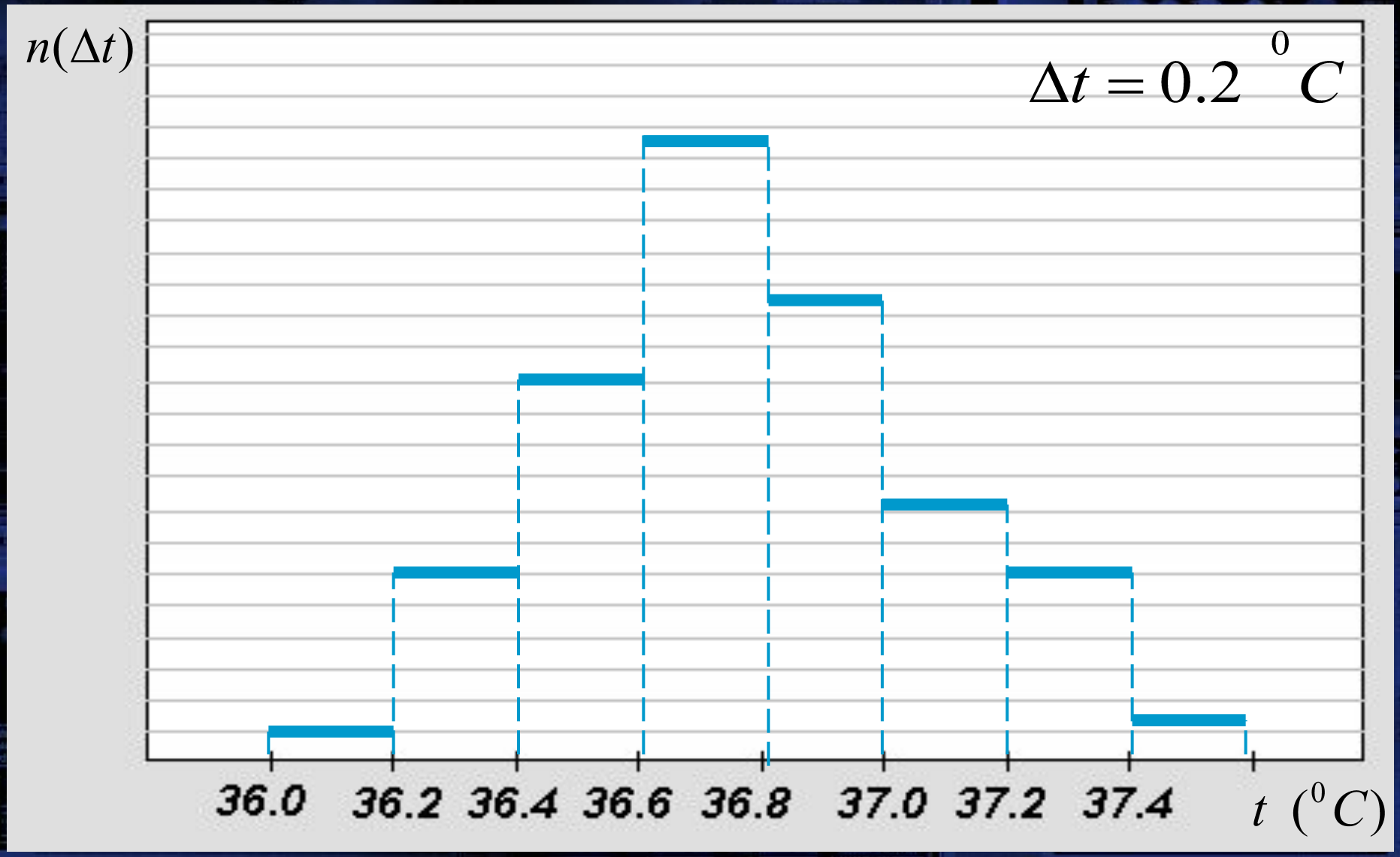
$$\overline{P}_i^N$$

$$\sigma_i^N$$

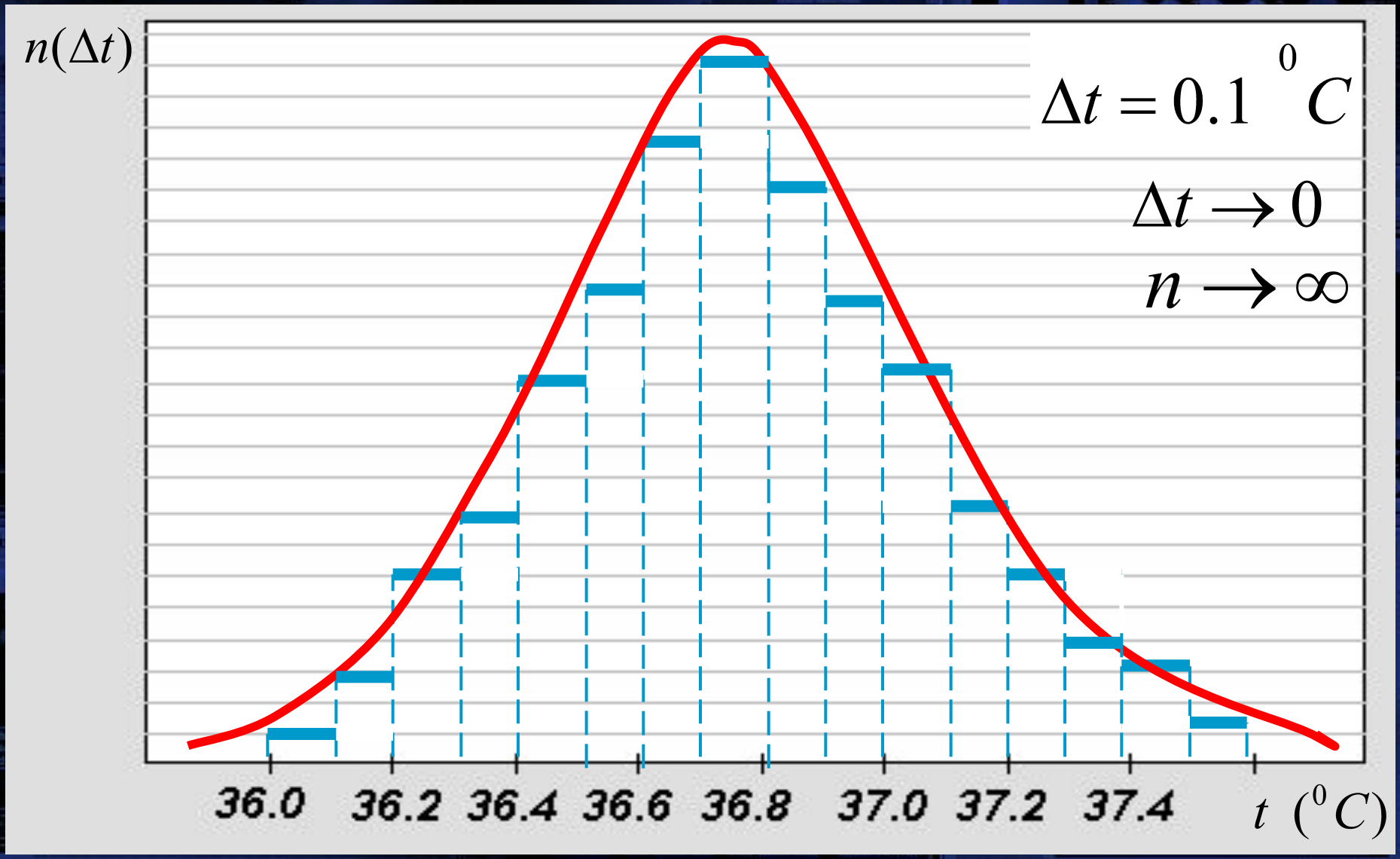
- Форма распределения; нормальность.

  Важным способом "описания" переменной является форма ее распределения, которая показывает, с какой частотой значения переменной попадают в определенные интервалы ее значений.
- Более точную информацию о форме распределения можно получить с помощью *критериев нормальности (Шапиро-Уилка).* Однако самым простым способом оценки распределения является построение гистограммы (графика, показывающего частоту попаданий значений переменной в отдельные интервалы).

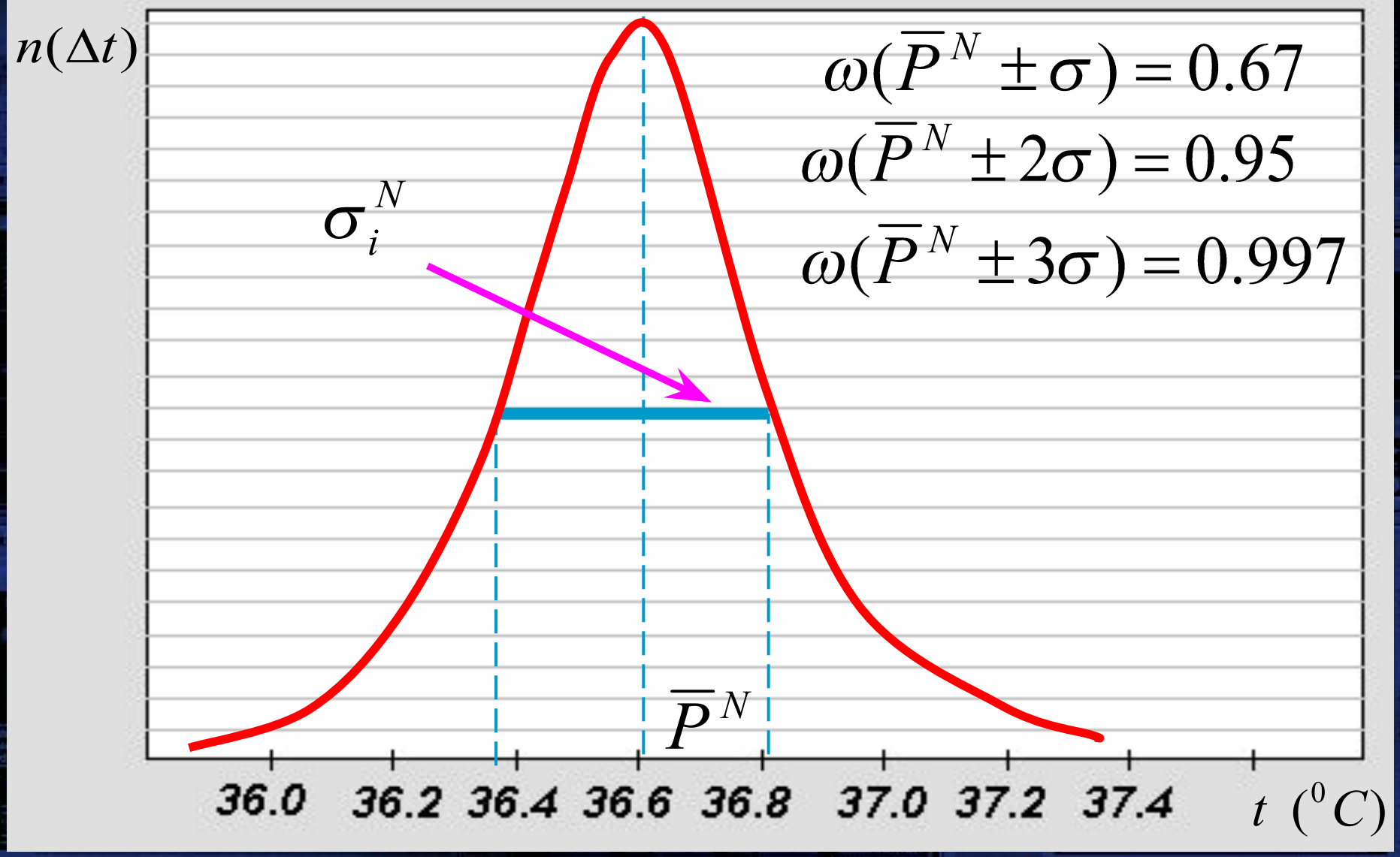
1. Определение вектора состояния в норме

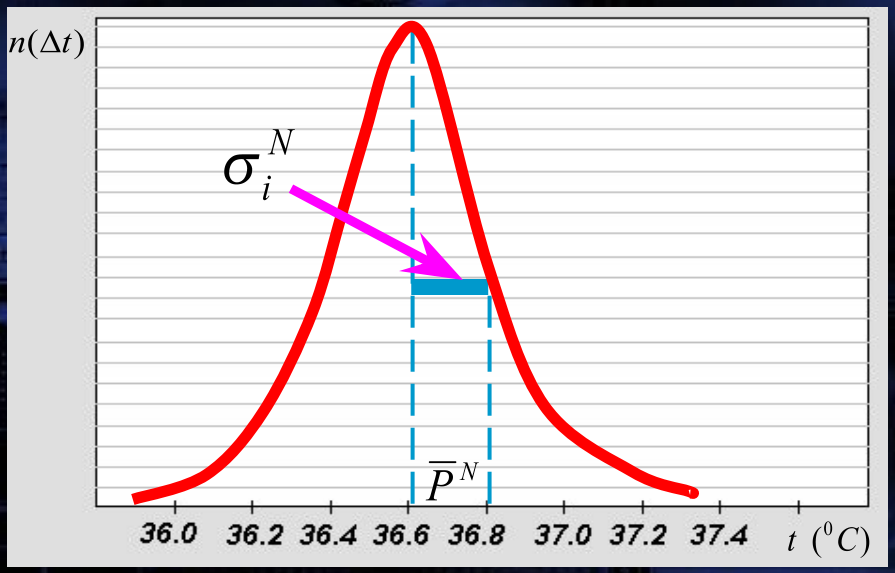


1. Определение вектора состояния в норме



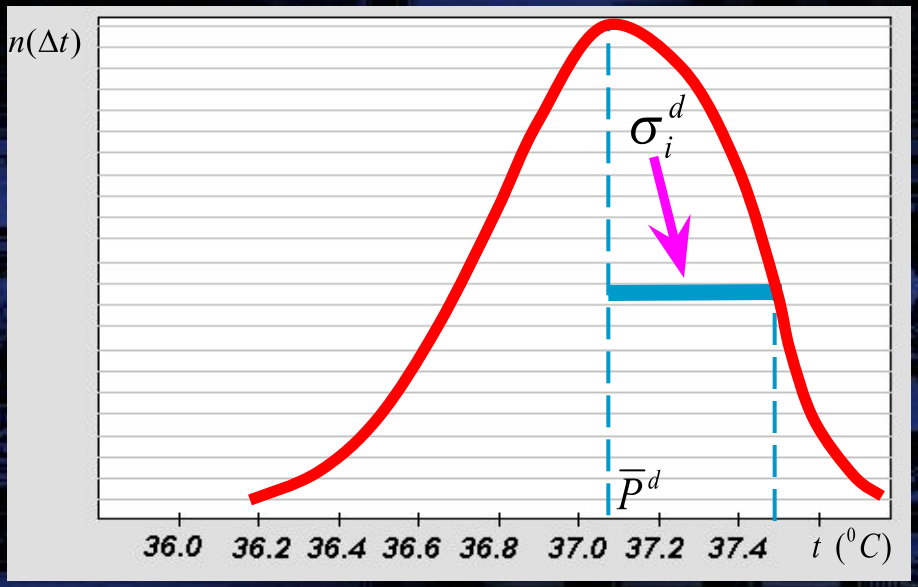
1. Определение вектора состояния в норме





1. Определение вектора состояния при заболевании

$\overline{P}_i^d \qquad \sigma_i^d$



Отличия:

1. $\overline{P}^N \neq \overline{P}^d$
2. $\sigma^d > \sigma^N$
3. **РАСПРЕДЕЛЕНИЯ НЕ СИММЕТРИЧНЫ**

- **Объем выборки.**

  Другим фактором, часто ограничивающим применимость критериев, основанных на предположении нормальности, является объем или размер выборки, доступной для анализа. До тех пор пока выборка достаточно большая (например, 100 или больше наблюдений), можно считать, что выборочное распределение нормально, даже если вы не уверены, что распределение переменной в популяции, действительно, является нормальным. Тем не менее, если выборка очень мала, то критерии, основанные на нормальности, следует использовать только при наличии уверенности, что переменная действительно имеет нормальное распределение.

Две основные задачи статистики
1.Нахождение различий выборок
2. Нахождение связи между выборками

Для нахождения различий между выборками распределенными нормально используются параметрические критерии (чаще t-критерий Стьюдента). Если же выборки малы и о их распределении ничего не известно используются непараметрические критерии. Говоря более специальным языком, непараметрические методы не основываются на расчетах параметров (таких как среднее или стандартное отклонение). Поэтому эти методы иногда также называются *свободными от параметров* или *свободно распределенными*.

- **Большие массивы данных и непараметрические методы.**

  Непараметрические методы наиболее приемлемы, когда объем выборок мал. Если данных много (например, $n > 100$), то не имеет смысла использовать непараметрические статистики. Главное здесь состоит в том, что когда выборки становятся очень большими, то выборочные средние подчиняются нормальному закону, даже если исходная переменная не является нормальной или измерена с погрешностью. Таким образом, параметрические методы, являющиеся более чувствительными (имеют большую статистическую мощность), всегда подходят для больших выборок.

Параметрический Т- критерий Стьюдента.

$$t_э = \frac{\left|\overline{P}_1 - \overline{P}_2\right|}{\sqrt{s_1^2 + s_2^2}}$$

где $S = \frac{\sigma}{\sqrt{n}}$

Эта величина сравнивается со значением коэффициента Стьюдента **t**, взятого из таблицы согласно объёму выборки n и заданной доверительной вероятностью **w** (или уровнем значимости Р). Если значение tэ превышает коэффициент **t**,

то с вероятностью **w** выборки, а следовательно и состояния различны.

Критические значения коэффициентов Стьюдента *t* для выборки объема *n* и заданной доверительной вероятности *ω*

	Доверительная вероятность ω (Уровень значимости p)	
	0,95 (0,05)	0,99 (0,01)
1	12,7	63,7
2	4,30	9,92
3	3,18	5,84
4	2,78	4,60
5	2,57	4,03
6	2,45	3,71
7	2,36	3,50
8	2,31	3,36
9	2,26	3,25
10	2,23	3,17
20	2,09	2,85
30	2,04	2,75
60	2,00	2,66

Критерий знаков (КЗ)

- **основан на подсчете однонаправленных эффектов в парных сравнениях;**
- **применяется для связанных (парных) выборок.**

Пример:

При измерении общего белка крови у 20 больных гепатитом было установлено, что у 17 больных этот параметр увеличился, а у 3 уменьшился по сравнению с нормой.

Необходимо установить, является ли повышение общего белка крови статистически значимым у больных гепатитом.

Решение:

Находим максимальное число менее часто встречающихся знаков изменения.

Максимальное число минусов = 3 (при общем числе опытов 20)

Сравниваем с табличным

Максимальное число знаков (менее часто встречающихся), при которых различия в парных сравнениях можно считать существенными

n	Уровень значимости p	
	0,05	0,01
6	0	0
7	0	0
8	1	0
9	1	0
10	1	0
11	2	1
12	2	1
13	3	1
14	3	2
15	3	2
20	5	4
30	10	8
40	14	12
50	18	16

Из таблицы видно, что для n=20 при p=0,05 допустимо 5 минусов. Пять больше чем три. Это значит, что повышение общего белка крови у больных гепатитом является статистически значимым.

Критерий Q Розенбаума

Основан на сравнении двух рядов наблюдений в общем упорядоченном ряду. Применяется для независимых выборок.

1. Подсчитывают число Q_1 и Q_2,

   где Q_1 – количество наблюдений первого ряда, которые больше максимальной величины второго ряда,

   Q_2 – количество наблюдений второго ряда, которые меньше минимальной величины первого ряда

2. Находят сумму $Q = Q_1 + Q_2$

ЕСЛИ ЧИСЛО НАБЛЮДЕНИЙ МЕНЬШЕ 11, КРИТЕРИЙ Q ПРИМЕНЯТЬ НЕЛЬЗЯ!

При любом числе наблюдений больше 26, различия можно считать существенными

для Qкр=8 при p=0.05

Критерий Q Розенбаума

Пример:

Сравнить max артериальное давление в мм. рт.ст. у детей с разными по тяжести угрожающими состояниями. Первая группа – дети с более легкими угрожающими состояниями, лечившиеся в отделениях общего типа. Вторая группа – дети с более тяжелыми угрожающими состояниями, лечившиеся в реанимационных отделениях и выздоровевшие.

	95;100;105;105;110;110;115;115;	120;130;135
75;80;80;85;90;	95;100;100;105;110;115;115;	

1. Q_1=3; Q_2=5,

   где Q_1 – количество наблюдений первого ряда, которые больше максимальной величины второго ряда,

   Q_2 – количество наблюдений второго ряда, которые меньше минимальной величины первого ряда

2. Находим сумму $Q = Q_1 + Q_2$=3+5=8
3. *Сравниваем найденное значение с табличным*

Критические значения Q-критерия Розенбаума. Минимальные значения Q, при которых различия между двумя выборками можно считать значимыми с вероятностью 95% (p=0,05)

n	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
p=0,05																
11	6															
12	6	6														
13	6	6	6													
14	7	7	6	6												
15	7	7	6	6	6											
16	7	7	7	7	6	6										
17	7	7	7	7	7	7	7									
18	7	7	7	7	7	7	7	7								
19	7	7	7	7	7	7	7	7	7							
20	7	7	7	7	7	7	7	7	7	7						
21	8	7	7	7	7	7	7	7	7	7	7					
22	8	7	7	7	7	7	7	7	7	7	7	7				
23	8	8	7	7	7	7	7	7	7	7	7	7	7			
24	8	8	8	8	8	8	8	8	8	8	7	7	7	7		
25	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7	
26	8	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7

Из таблицы видно, что для p=0.05, n_1=11 и n_2=12 Qкр=6.

Q=8 больше Qкр=6

Следовательно, различия существенны.

Корреляционный и регрессионный анализ

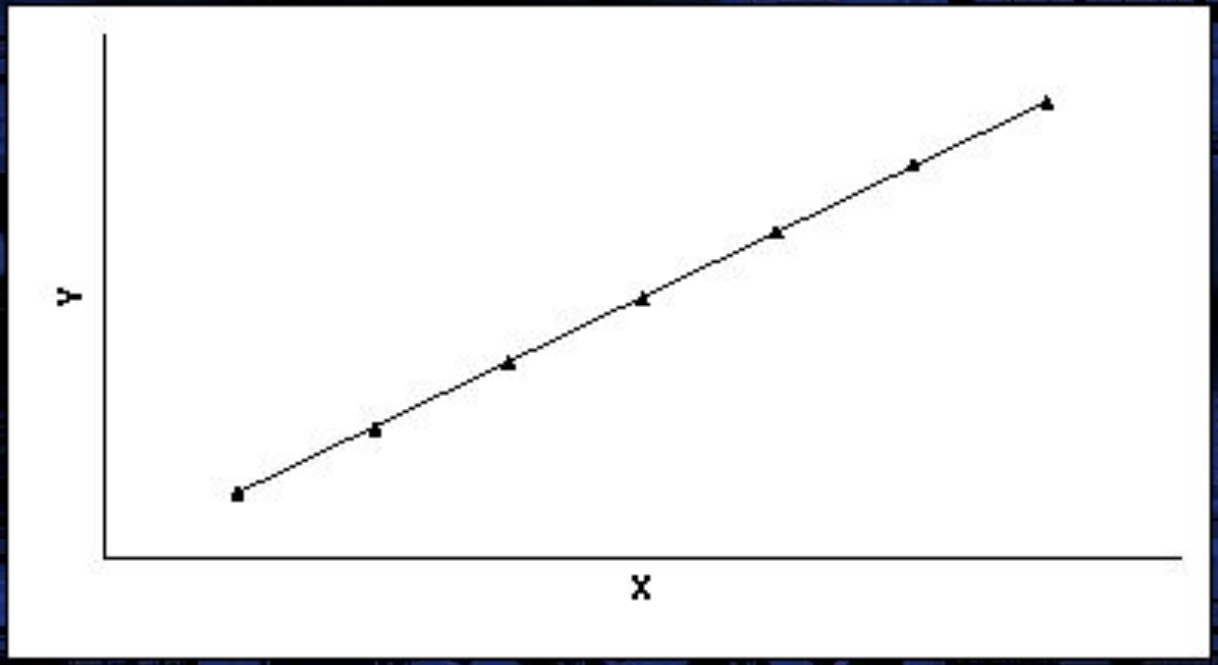
- **связь как *синхронность* (согласованность) – корреляционный анализ.**
- **связь как *зависимость* (влияние) – регрессионный анализ.**

Этапы анализа

1. **выявление наличия взаимосвязи между параметрами;**
2. **определение формы связи;**

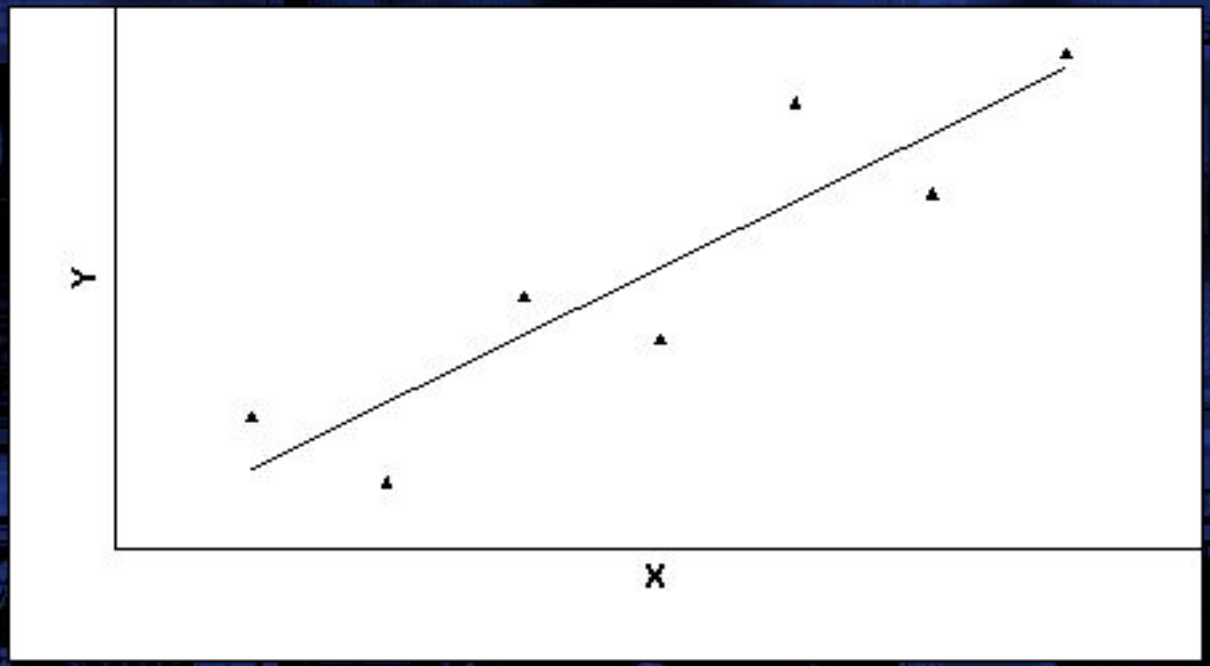
1. Выявление наличия связи между параметрами

Пример положительной функциональной связи между параметрами X и Y.

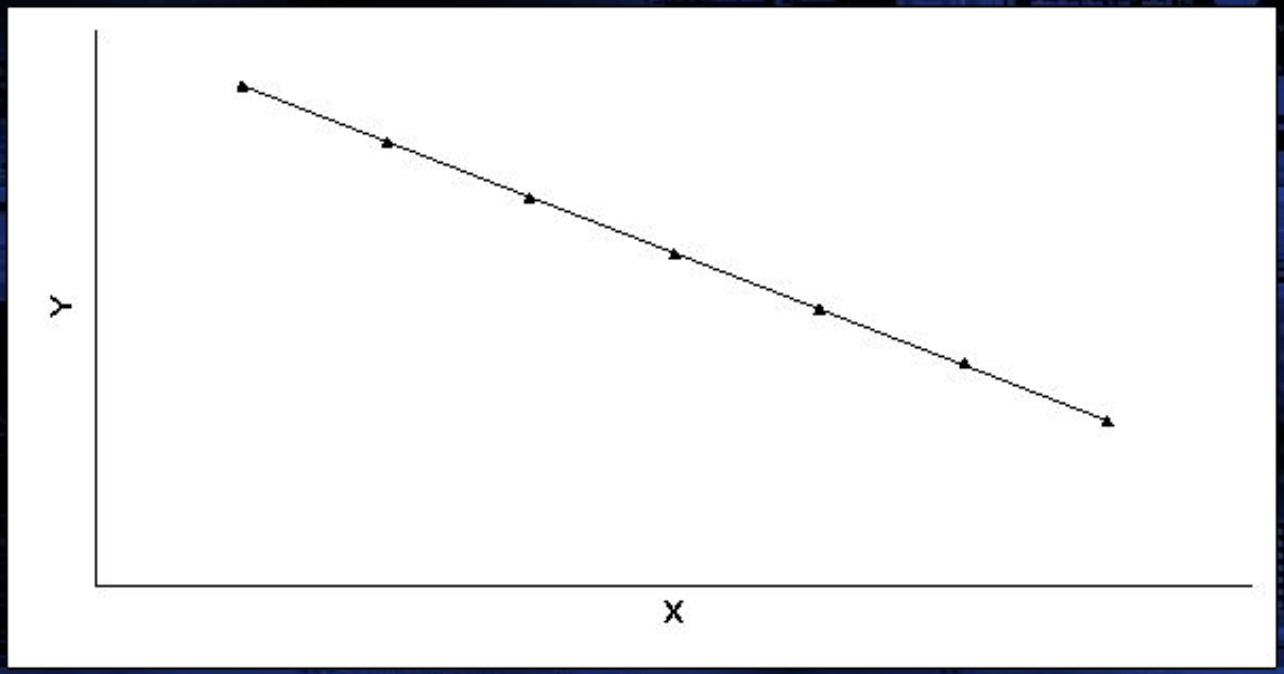


- Чем больше значения одного параметра, тем больше значения другого.

Пример положительной статистической связи между параметрами X и Y.

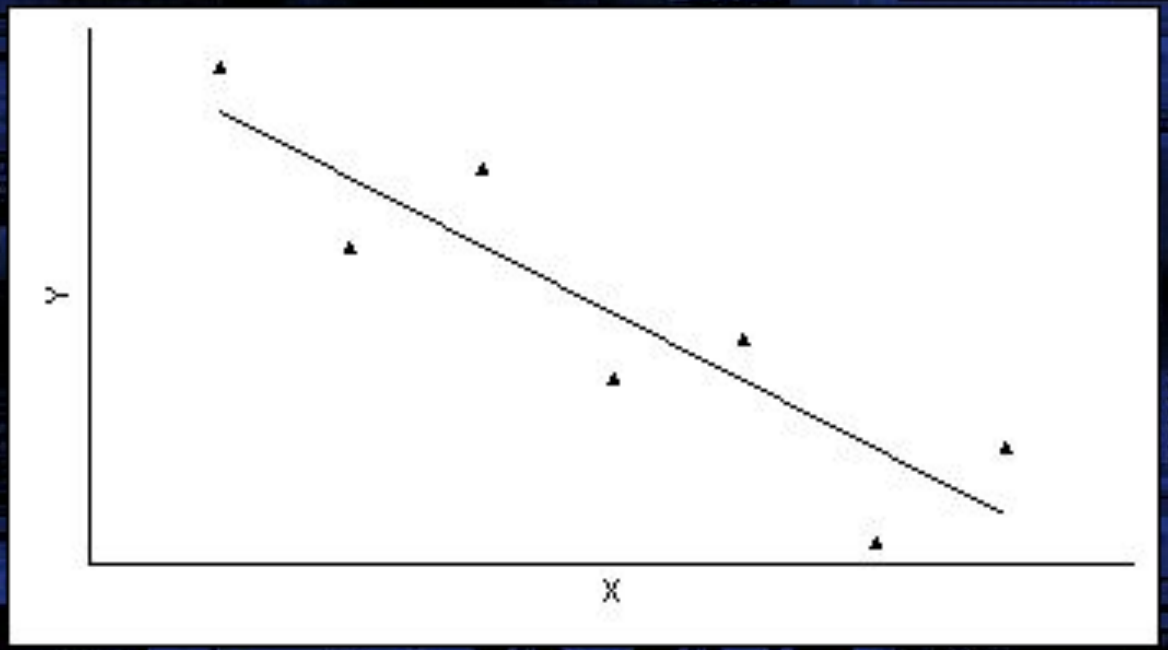


Пример <u>отрицательной функциональной связи</u> между параметрами X и Y.



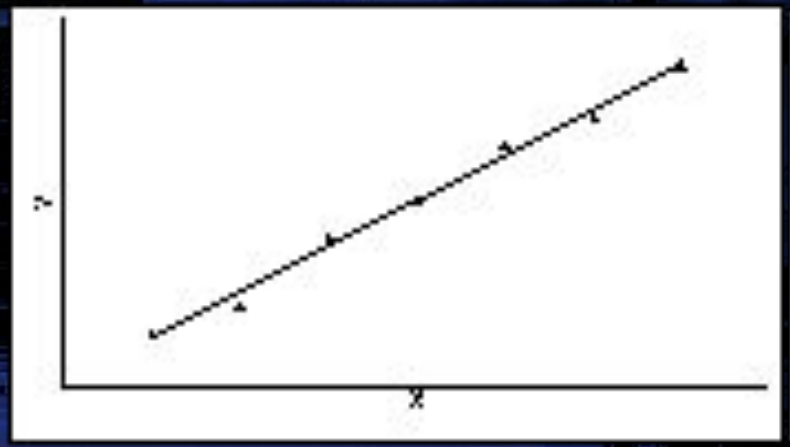
- *Чем больше* значения одного параметра, *тем меньше* значения другого.

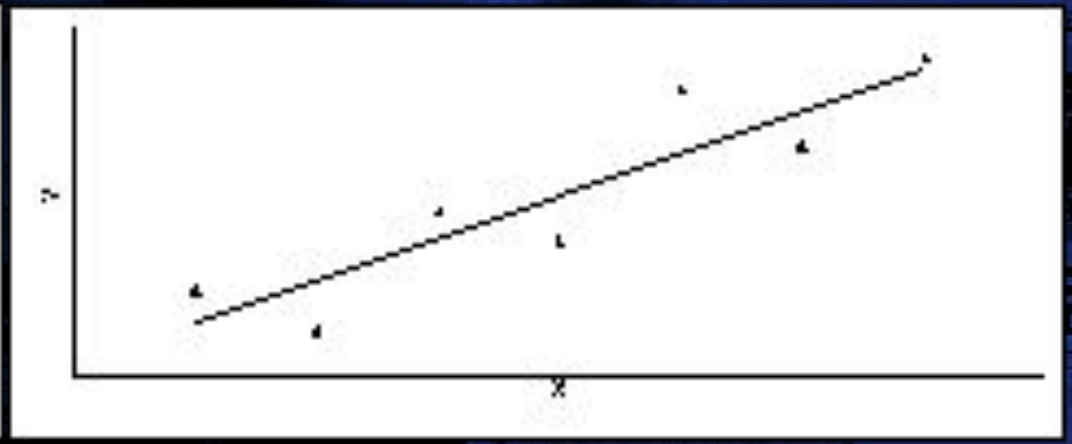
Пример отрицательной статистической связи между параметрами X и Y.



Определение силы (тесноты) связи

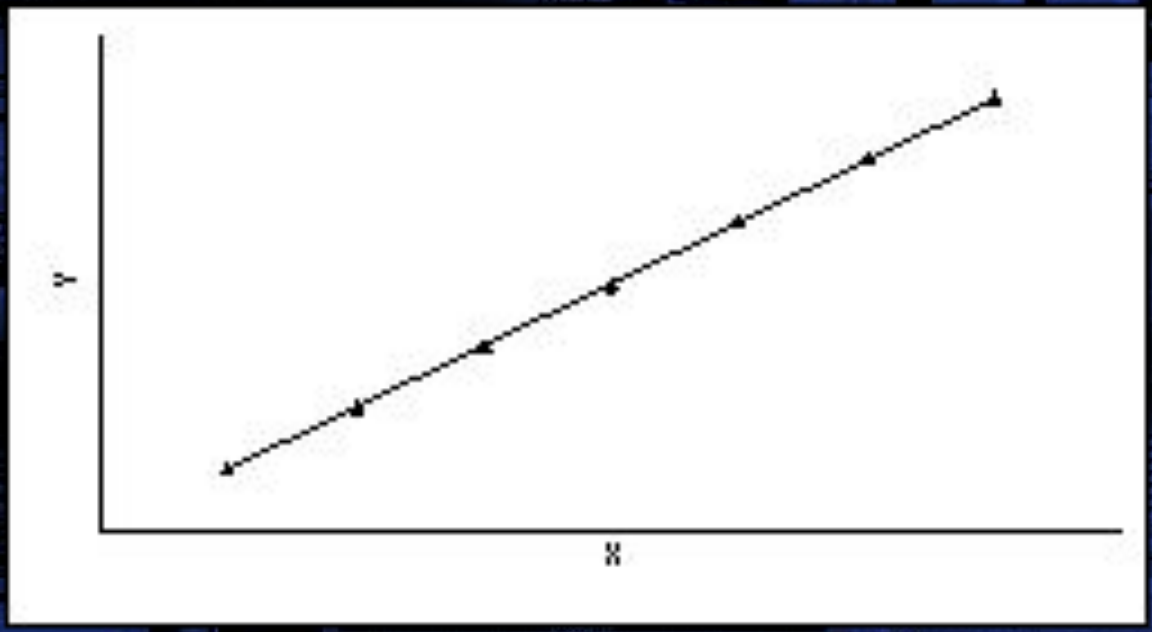
Коэффициент парной корреляции показывает, насколько тесно две переменные связаны между собой.





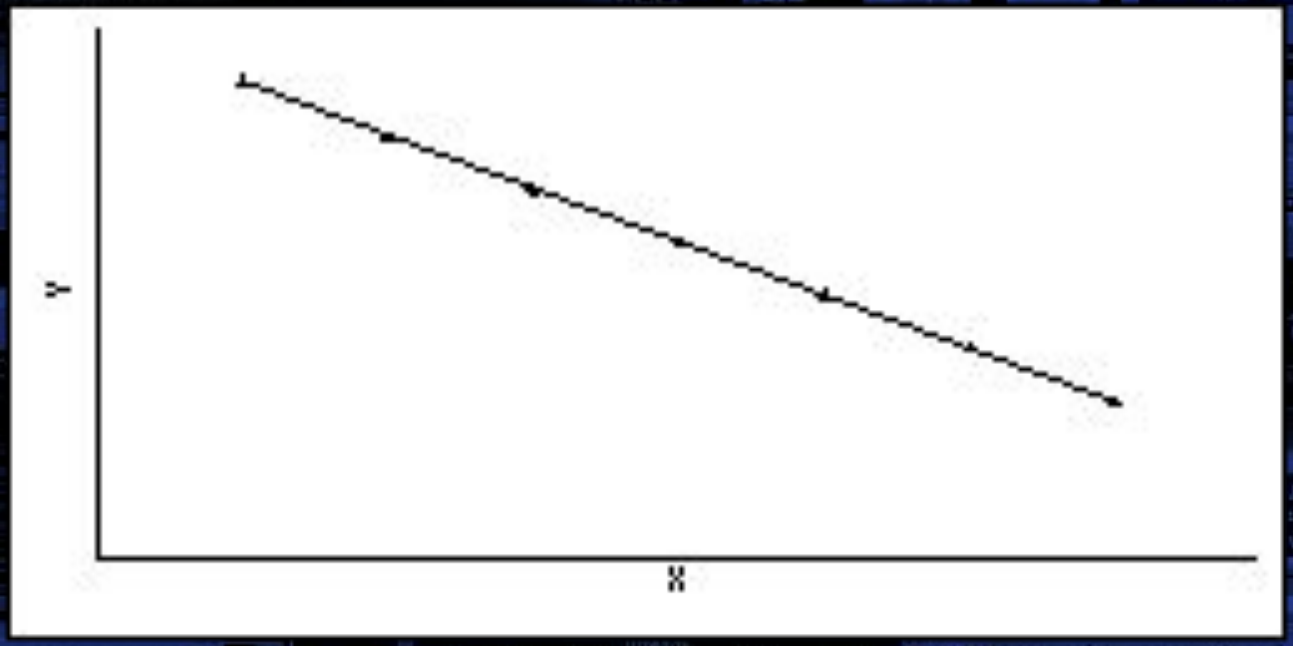
Коэффициент парной корреляции *r* принимает значения в диапазоне от –1 до +1.

Коэффициент корреляции
Мера тесноты линейной связи



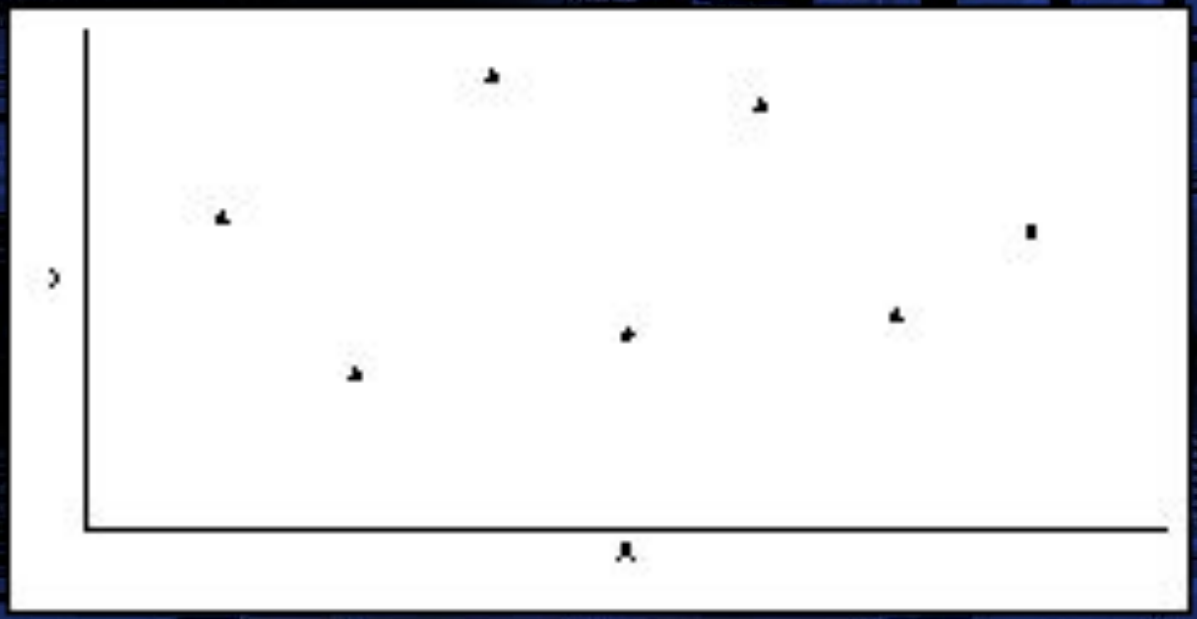
Если ***r = 1***, то между двумя переменными существует ***функциональная положительная линейная связь***, т.е. на диаграмме рассеяния соответствующие точки лежат на одной прямой с положительным наклоном.

Коэффициент корреляции
Мера тесноты линейной связи



Если ***r = -1***, то между двумя переменными существует ***функциональная отрицательная линейная зависимость***, т.е. на диаграмме рассеяния соответствующие точки лежат на одной прямой с отрицательным наклоном.

Коэффициент корреляции
Мера тесноты линейной связи



Если ***r = 0***, то рассматриваемые ***переменные линейно независимы***, т.е. на диаграмме рассеяния облако точек "вытянуто по горизонтали".

Коэффициент корреляции

Формула для вычисления парного коэффициента линейной корреляции:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- вычисляется для количественных признаков;
- симметричен;
- величина *безразмерная*;
- *не изменяется при изменении единиц измерения* параметров X и Y;
- **d=r²** – *коэффициентом детерминации* (выражается в %)

d – это показатель того, насколько изменения зависимого признака объясняются изменениями независимого (сила связи).

- Коэффициент детерминации принимает значения в диапазоне от 0% до 100%.

Коэффициент корреляции и детерминации

- если две переменные линейно независимы (метод наименьших квадратов дает горизонтальную прямую), то одна из них в своих изменениях никоим образом не определяет другую, $d = 0$.
- коэффициент детерминации указывает, какая часть изменений одной переменной объясняется изменениями другой переменной.
- ***чем выше*** по модулю (по абсолютной величине) значение коэффициента корреляции, ***тем сильнее связь*** между параметрами.

Принято считать, что если

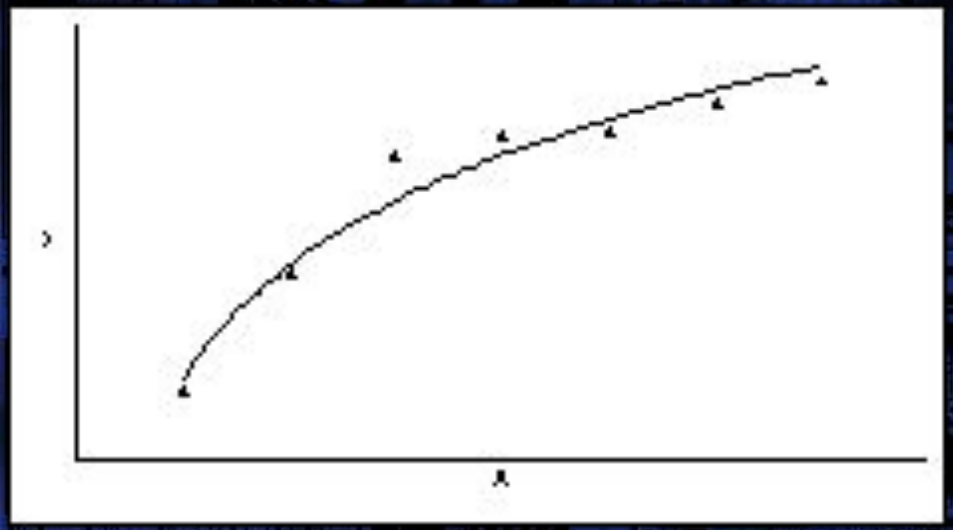
$|r| > 0.7$ - **сильная корреляционная связь** (при этом $d > 50\%$, т.е. один параметр определяет другой более, чем наполовину).

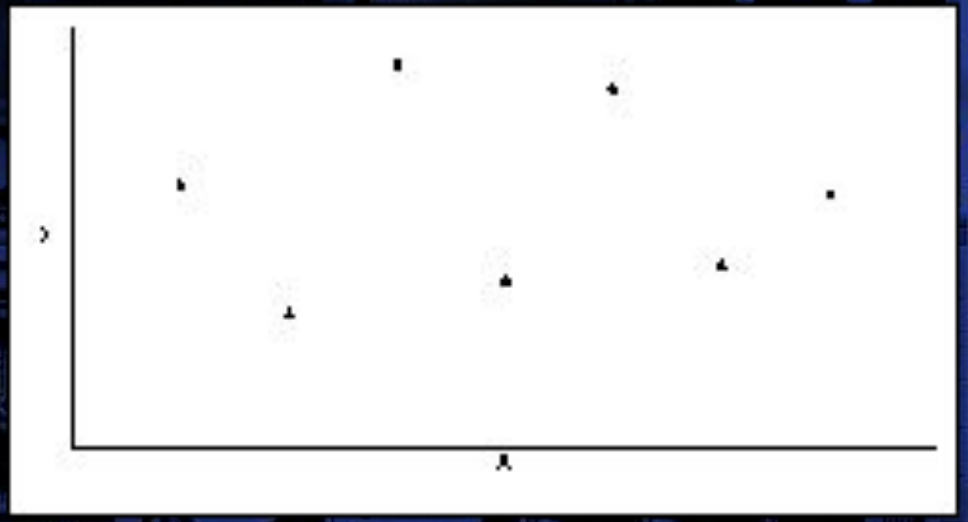
$0.3 < |r| < 0.7$ - **связь средней силы** (при этом $10\% < d < 50\%$)

$|r| < 0.3$ - **слабая связь** (при этом $d < 10\%$).

2. Подбор формы связи

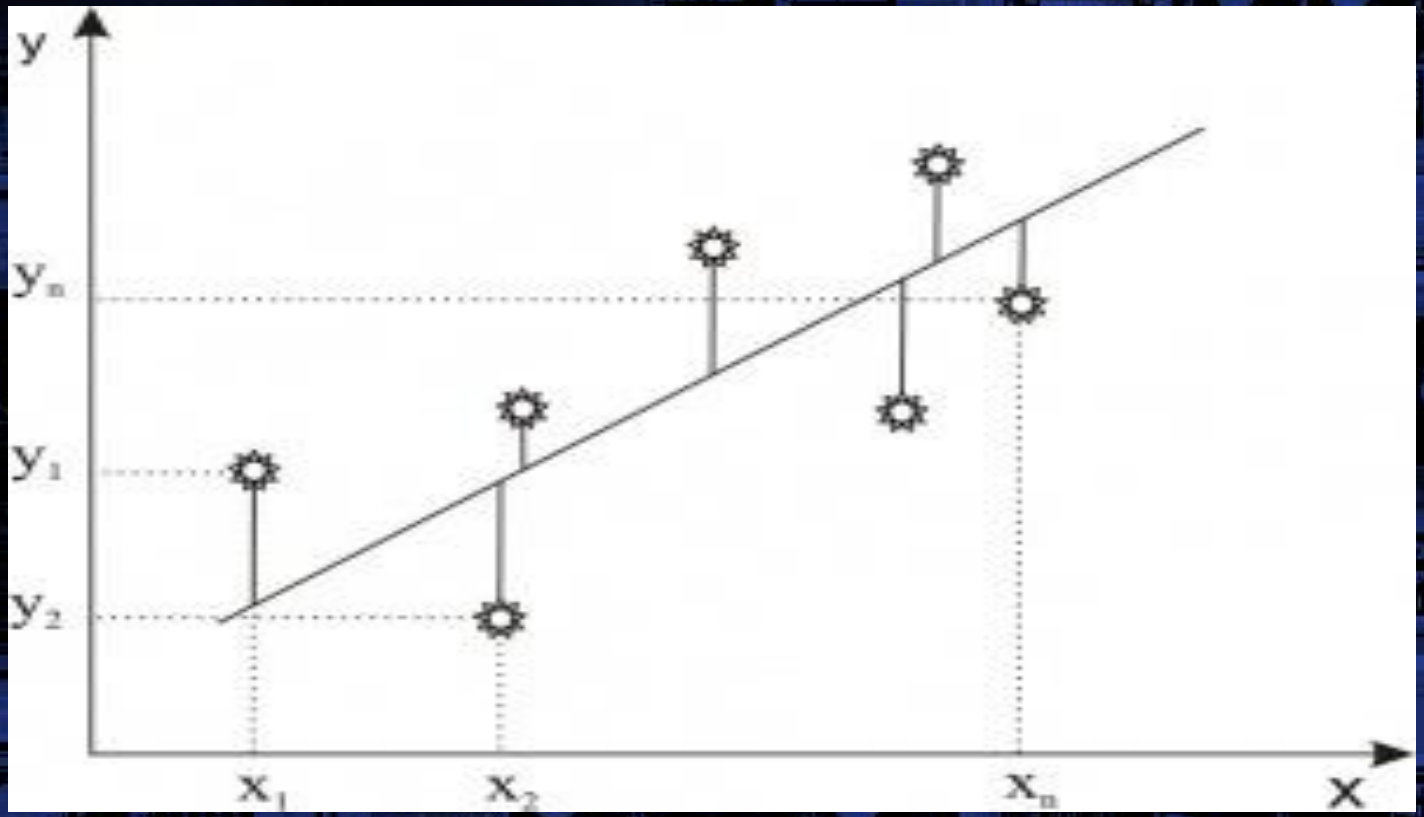
- Линейная и нелинейная связь
- Отсутствие связи между параметрами





2. Подбор формы связи

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ



ЛИНИЯ РЕГРЕССИИ.

Сумма квадратов расстояний от точек на диаграмме до этой линии минимальна (по сравнению со всеми возможными линиями).

Коэффициенты аппроксимирующих формул

Если связь есть, то ее можно описать с помощью аппроксимирующей формулы.

Вводим данные в компьютер и рассчитываем выборочный и начальный коэффициенты регрессии (*a* и *b*).

$$P_2 = aP_1 + b$$

– **ЛИНЕЙНАЯ ЗАВИСИМОСТЬ**

Если распределение не является нормальным, то можно перейти к непараметрическим коэффициентам корреляции, одинаково пригодным при любом непрерывном распределении.

Для расчета непараметрического *коэффициента ранговой корреляции Спирмена* необходимо сделать следующее. Для каждого *xi* рассчитать его ранг *ri* в вариационном ряду, построенном по выборке X. Для каждого *yi* рассчитать его ранг *qi* в вариационном ряду, построенном по выборке Y. Для набора из *n* пар вычислить линейный коэффициент корреляции. Он называется коэффициентом ранговой корреляции, поскольку определяется через ранги.

В качестве примера рассмотрим данные роста и веса десяти марсиан из книги С. Гланца:

№	1	2	3	4	5	6	7	8	9	10
Рост(см)	33	35	31	40	42	32	41	34	35	46
Вес(г)	7,6	9,6	7,7	11,8	14,8	7,7	12,2	9,1	9,9	15,6

Рост(см)	Ранг (ri)	Вес(г)	Ранг (qi)	ri - qi	$(ri - qi)^2$
31	1	7,7	2	-1	1
32	2	8,3	3	-1	1
33	3	7,6	1	2	4
34	4	9,1	4	0	0
35	5,5	9,6	5	0,5	0,25
35	5,5	9,9	6	-0,5	0,25
40	7	11,8	7	0	0
41	8	12,2	8	0	0
42	9	14,8	9	0	0
46	10	15,6	10	0	0
				Сумма	6,5

Формула для расчета коэффициента ранговой корреляции Спирмена

$$\rho_n = 1 - \frac{6\sum_{i=1}^{n}(r_i - q_i)^2}{n^3 - n}.$$

Используя данную формулу вычислим коэффициент ранговой корреляции Спирмена

$$\rho_s = 1 - \frac{6*6,5}{1000-10} = 0,96$$

Обратимся к таблице критических значений коэффициента ранговой корреляции Спирмена

n	Уровень значимости p	
	0,05	0,01
6	0,886	1,0
7	0,786	0,929
8	0,738	0,881
9	0.70	0,833
10	0,648	0,794
15	0.521	0.654
20	0,447	0,570
25	0.398	0,511
30	0,362	0,467

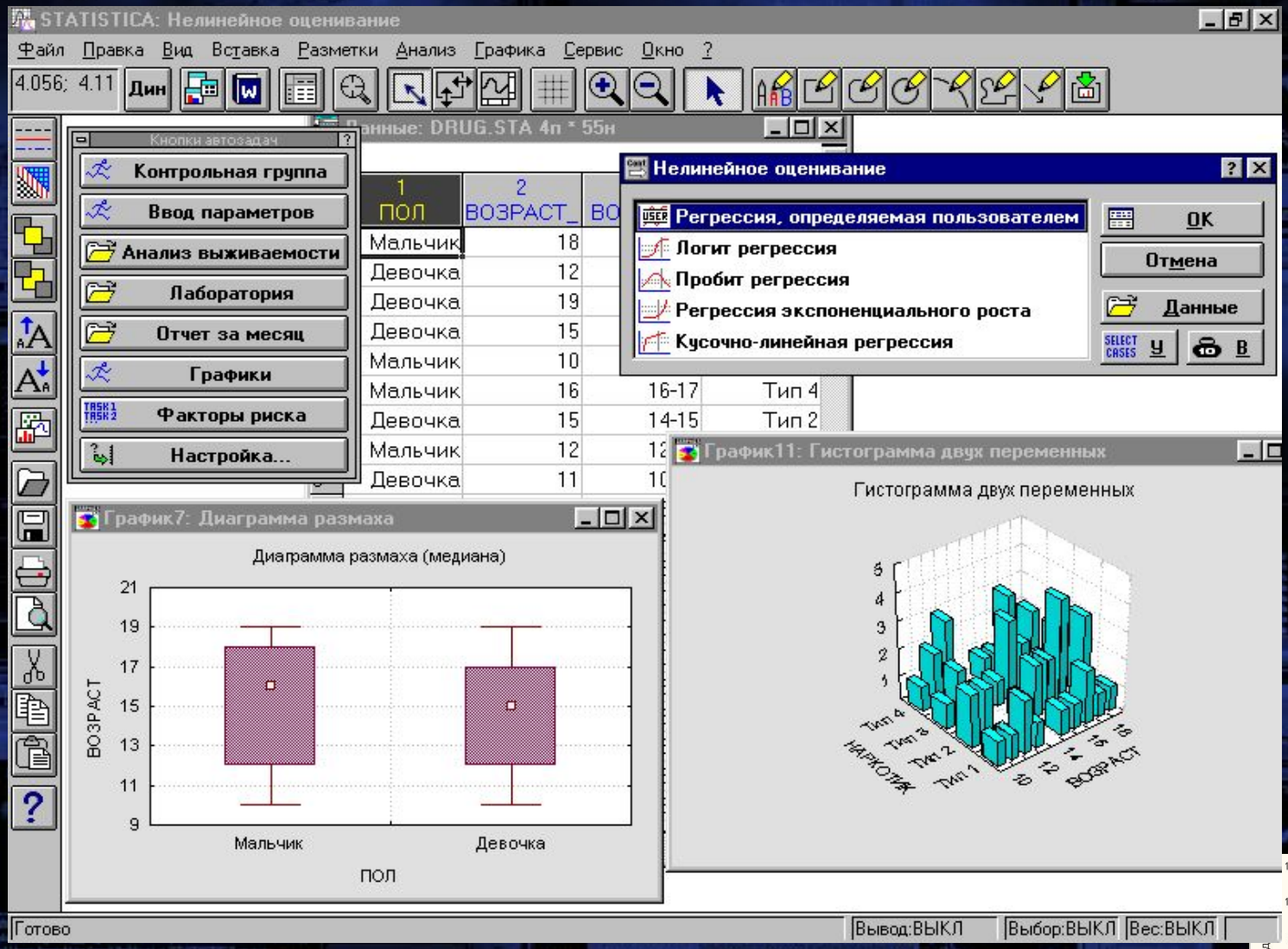
Критическое значение для уровня значимости 0,01 и объема выборки n=10 равно 0,794, что меньше полученного нами (0,96). Т.е. корреляция статистически значима (P<0,01).

Пакеты программ для статистической обработки медицинской и биологической информации

О современных системах статистического анализа на персональных компьютерах

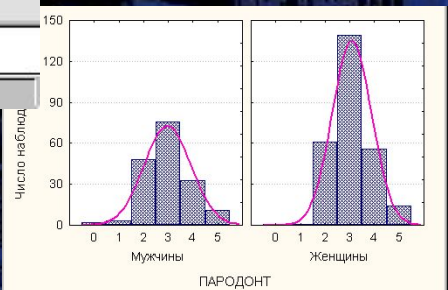
- ***STATISTICA***
- ***SPSS***
- ***S-плюс***
- ***SAS***
- ***MStat***

Система *STATISTICA*

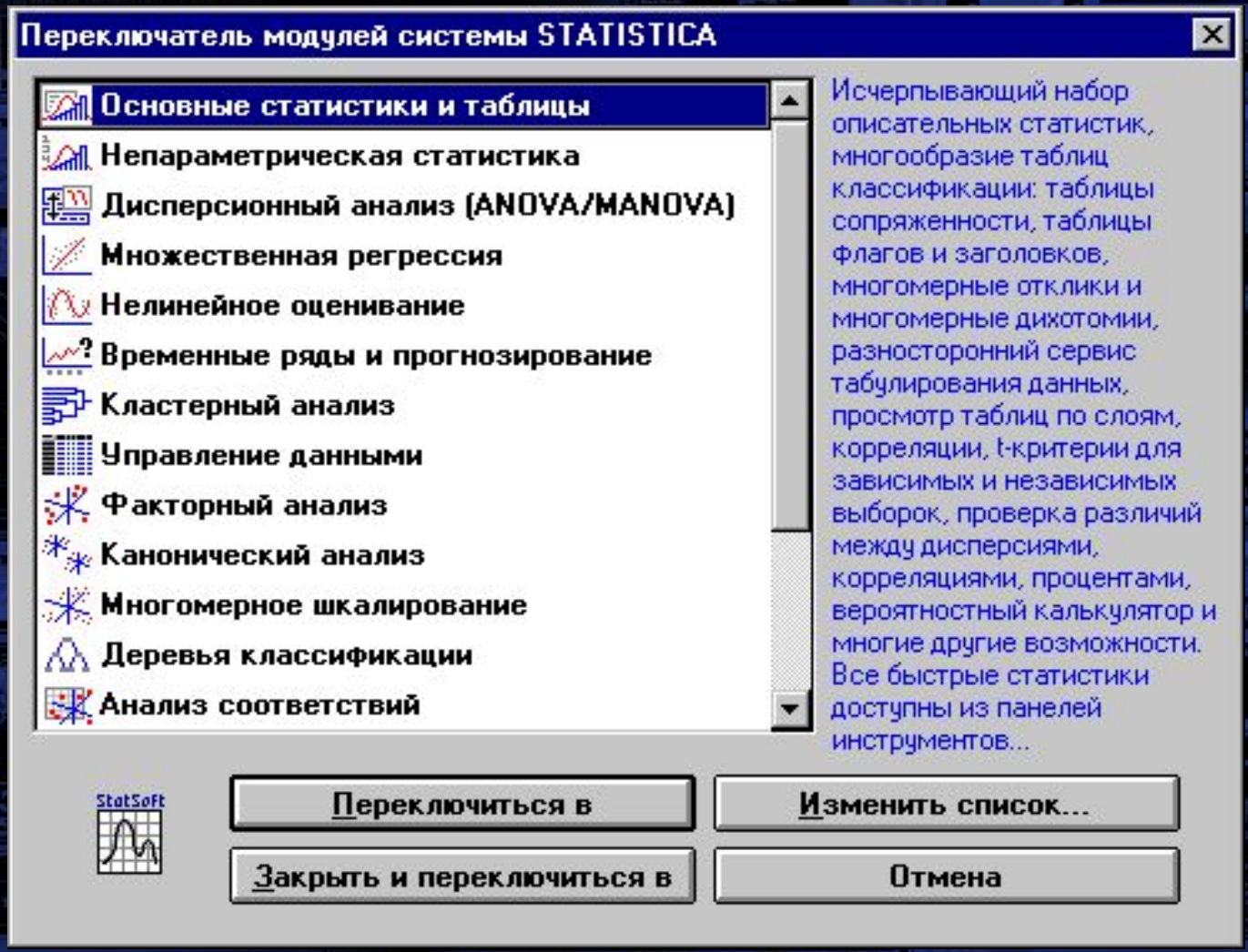


StatSoft

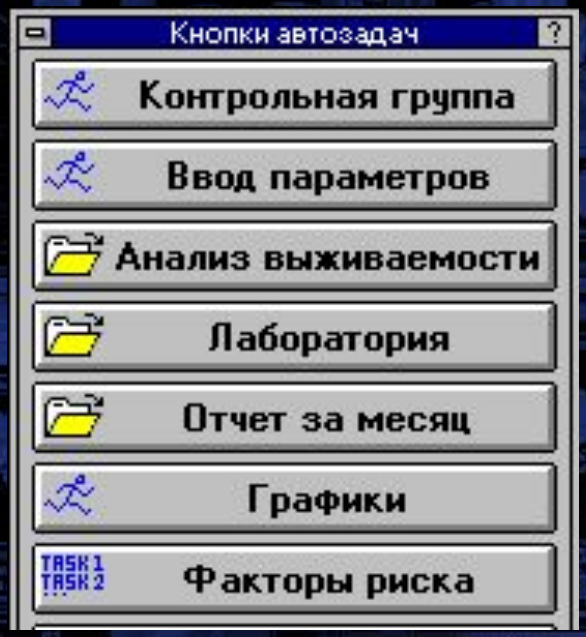
Реализован графически-ориентированный подход к анализу данных

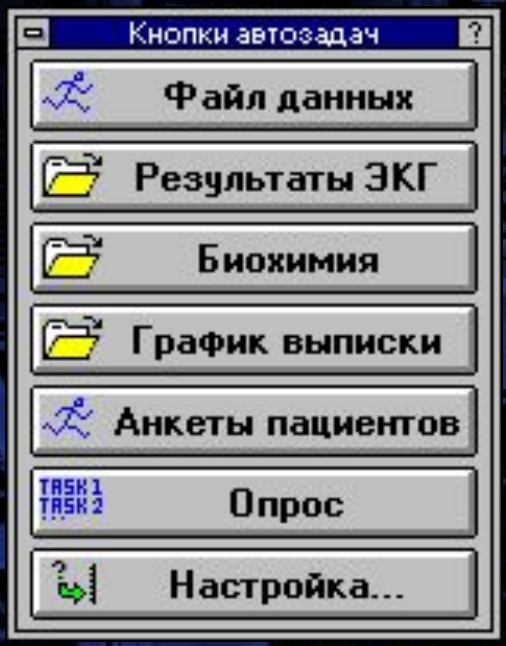


Система *STATISTICA* состоит из отдельных модулей, покрывающих весь спектр современных методов анализа данных.



Гибкий интерфейс, отвечающий всем стандартам Windows, позволяет настроить систему под конкретный проект, связанный с анализом медицинских данных.





Основные этапы анализа данных

Подготовка данных: заполнение таблиц, импорт, проверка и сортировка.

Разведочный анализ: основные статистики и графики.

Анализ зависимостей.

Построение моделей.

Типы медицинской информации

- Массовые обследования (десятки тысяч наблюдений и сотни показателей).
- Результаты клинических исследований (наблюдения за группами пациентов).

Количественные и качественные признаки. Группирующие переменные.

Переменные

Наблюдения

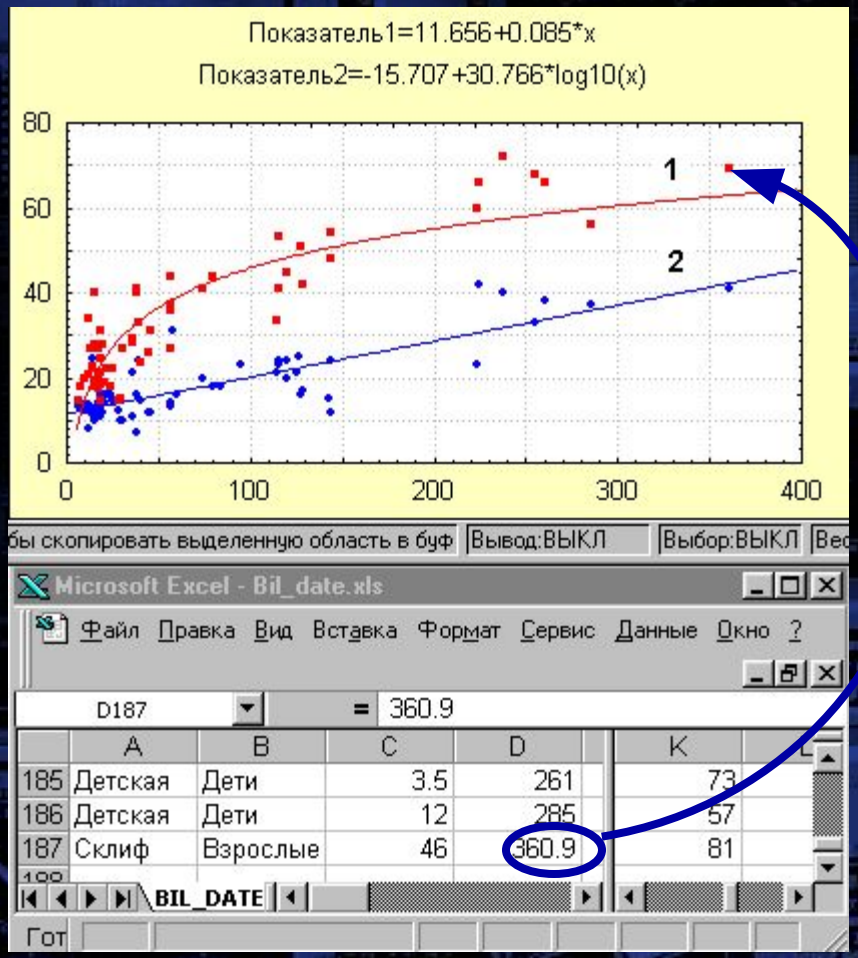
Данные: BIL_DATE.STA 17п * 186н

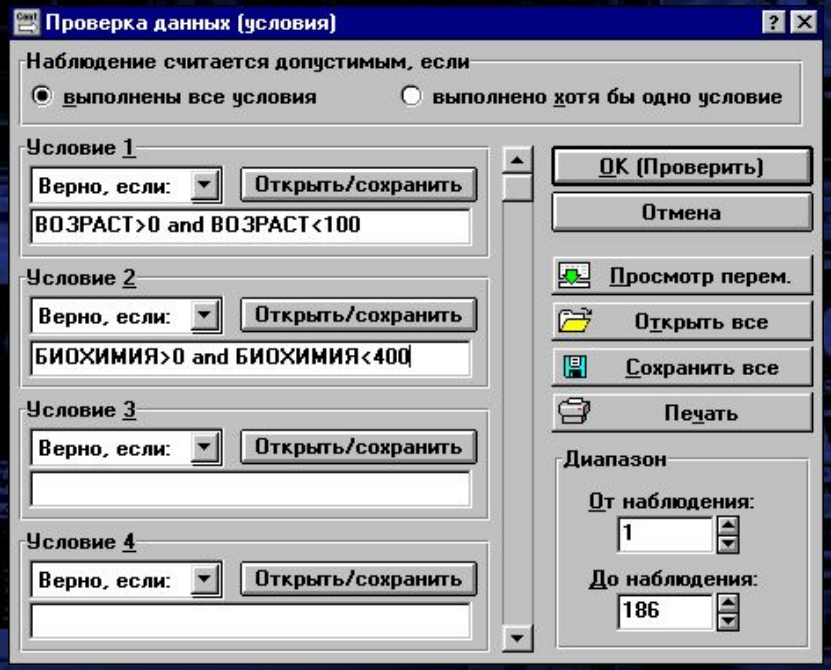
ТЕКСТОВЫЕ ЗНАЧЕНИЯ: Неинвазивные измерения билирубина в сравнении с биохимическим анализом

	1 МЕСТО	2 ГРУППА	3 ВОЗРАСТ	4 БИОХИМИЯ	9 ПЛЕЧО	10 ЛОБ
130	Склиф	Взрослые	77.0	35.5	28.2	28.4
131	Детская	Дети	8.0	35.5	29.0	21.0
132	Детская	Дети	14.0	38.5	40.0	22.0
133	Детская	Дети	9.0	38.5	41.0	26.0
134	Детская	Дети	9.0	39.0	33.0	32.0
135	Склиф	Взрослые	54.0	40.2	23.6	25.0
136	Детская	Дети	10.0	43.5		
137	Склиф	Взрослые	22.0	44.8	26.0	24.6
138			12.0	45.5	31.0	23.0
139			7.0	45.5		
140			12.0	45.5	31.0	23.0
141	Детская	Дети	7.0	45.5		
142	Новорожд	Новорожд	0.0	53.0		54.0
143	Детская	Дети	2.5	54.0		
144	Детская	Дети	10.0	57.0	37.0	28.0
145	Детская	Дети	7.0	57.0	27.0	18.0
146	Детская	Дети	14.0	57.0	44.0	29.0
147	Детская	Дети	12.0	57.0	36.0	36.0
148	Инф.б.на	Взрослые	20.0	57.0		30.0

Подготовка информации

Импорт из баз данных, текстовых файлов или электронных таблиц.

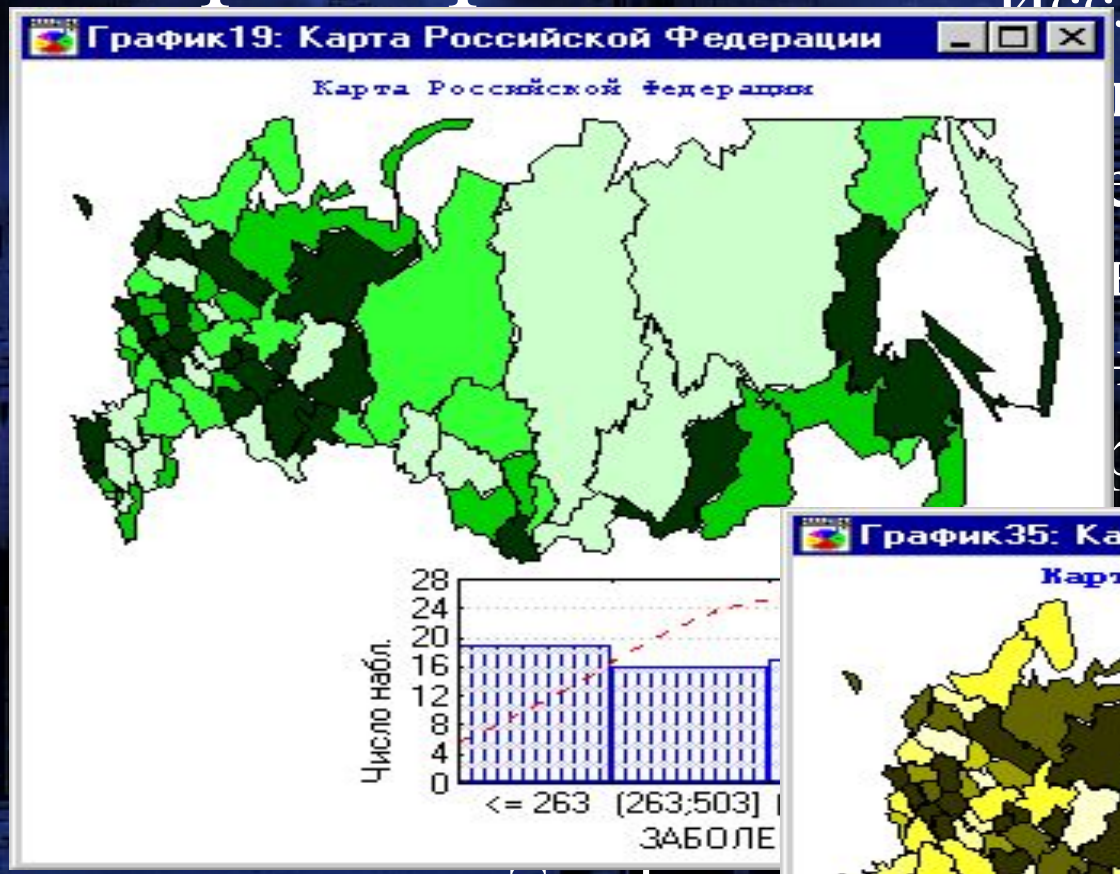


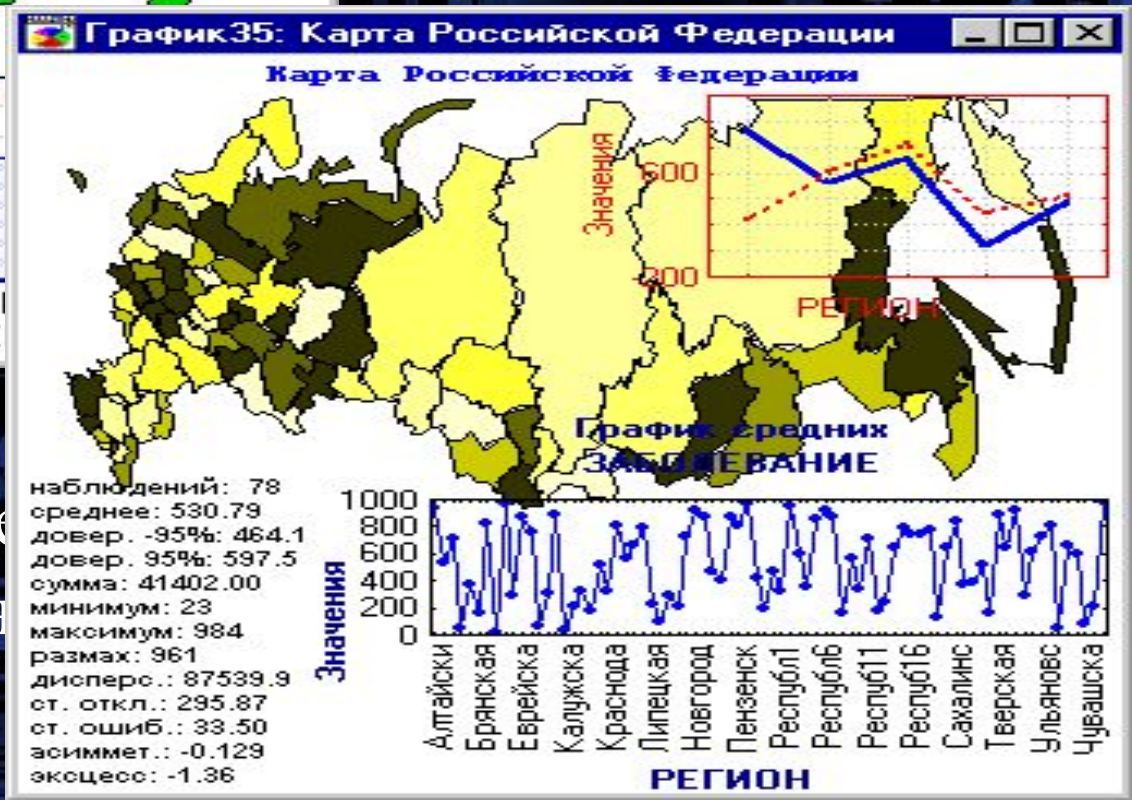


Динамический обмен данными (DDE) с исходным файлом.

Пример:

Исследуется прибор для
нвазивного измерения
ержания билирубина в
ви. Измерения в
ных точках тела
релируют с данными





групп пациентов.
Возможно построение
модели после разбиен
данных на группы.

Система *STATISTICA*

www.statsoft.ru

Учебник содержит разделы по методам статистического анализа данных и предназначен в первую очередь для тех, кто не является специалистом по математической статистике.


СПАСИБО ЗА ВНИМАНИЕ!!

Система *MSTAT*