Регрессионный анализ

Методы анализа	Факторы Независи-мые	Отклики Зависимые	Результат
Дисперсионный	Любые шкалы	Интервальные	Стат.значимость и сила влияния
Корреляционный	Нет разделения Любые шкалы (разные коэффициенты)		Сила и направление связи
Регрессионный	Интервальные (предикторы)	Интервальные	Прогноз (интер- но не экстаполяция ?)

Регрессия

- Моделирование, описание зависимости между переменными
- Количественная оценка поведения отклика при изменении предиктора

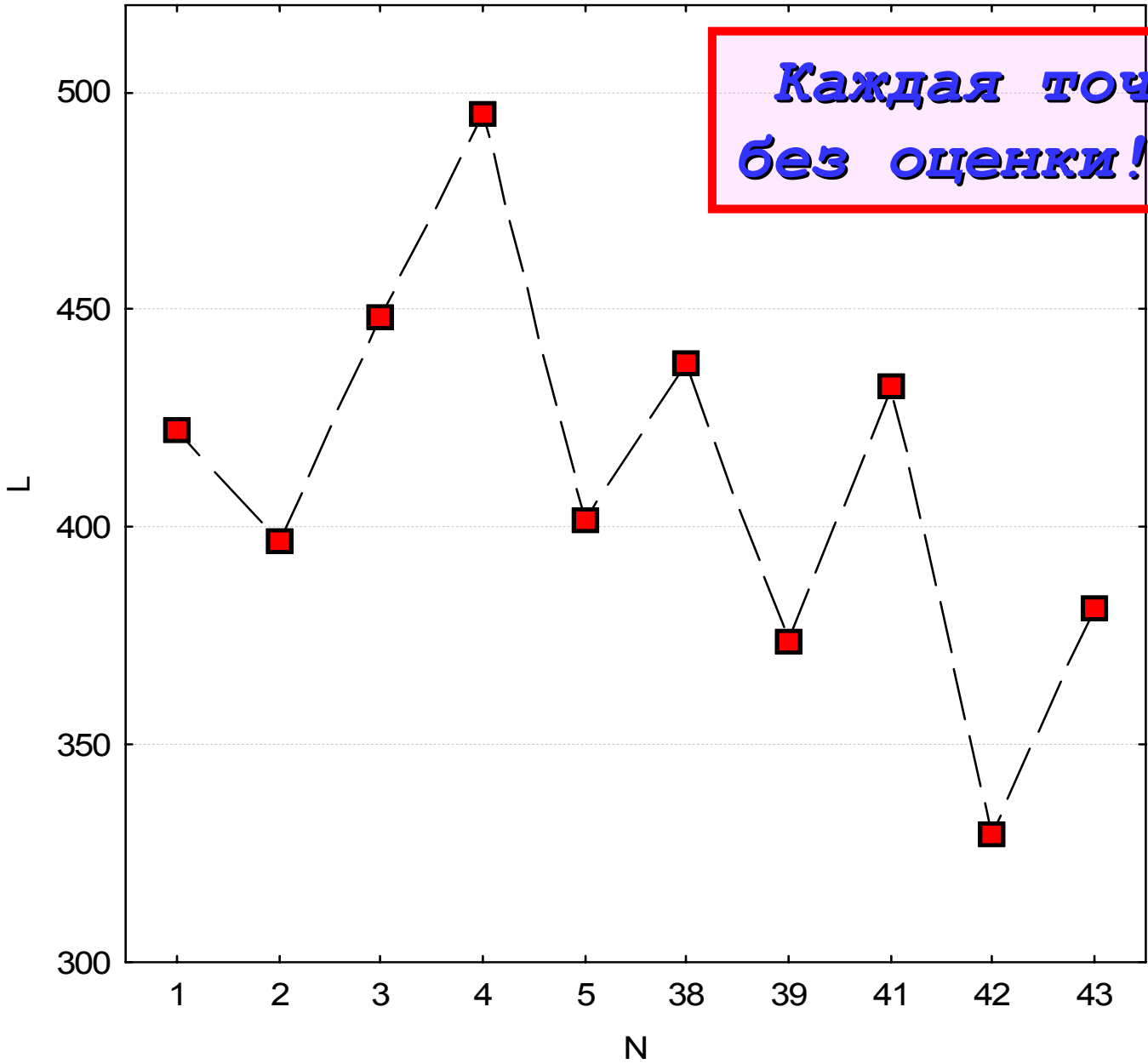
  - >> *уравнение регрессии*
- Предсказание значений переменной отклика при заданных значениях предиктора

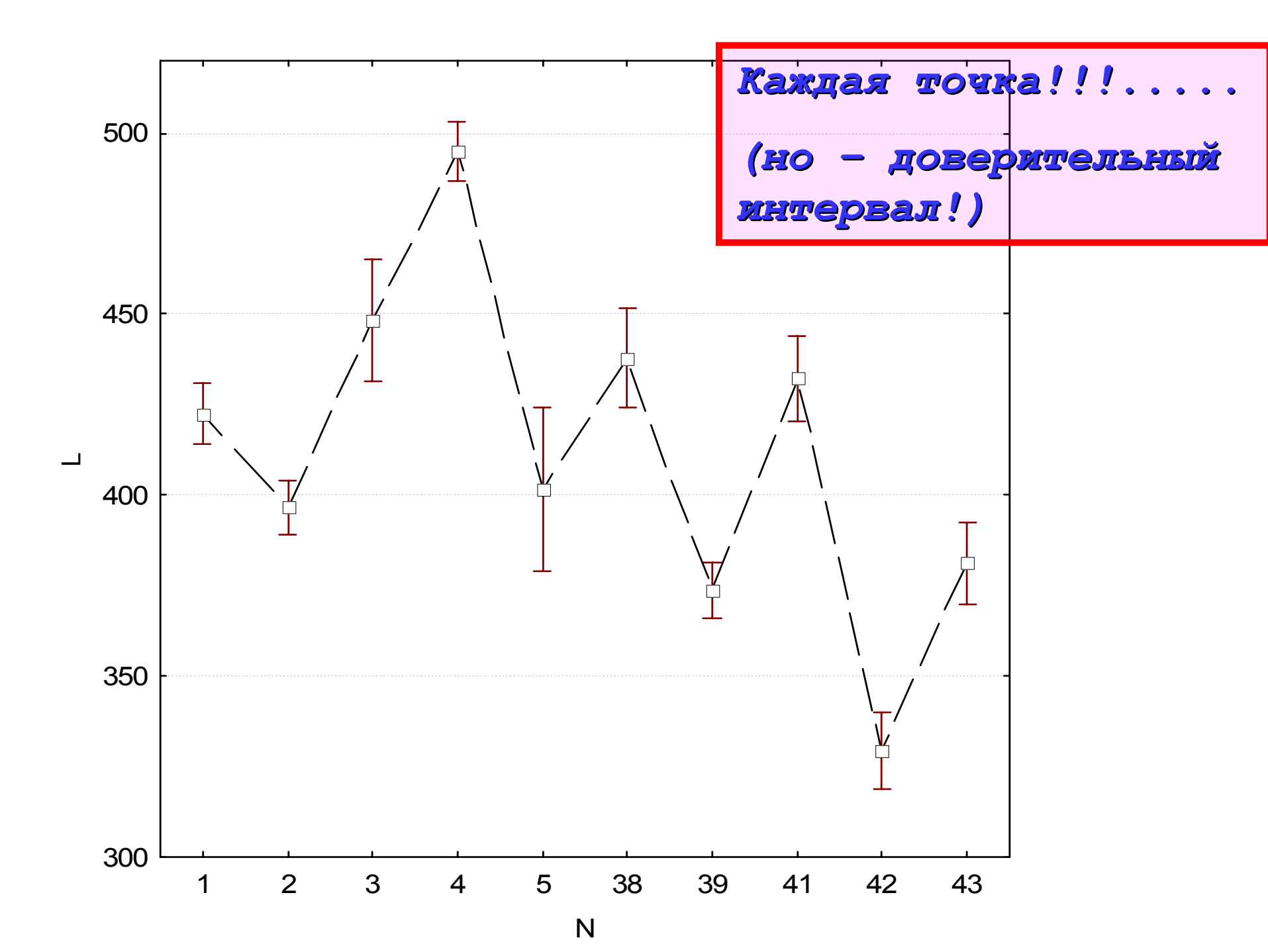
  - >> *прогноз*

Довольно часто для описания зависимости достаточно получить графическое изображение имеющихся данных.

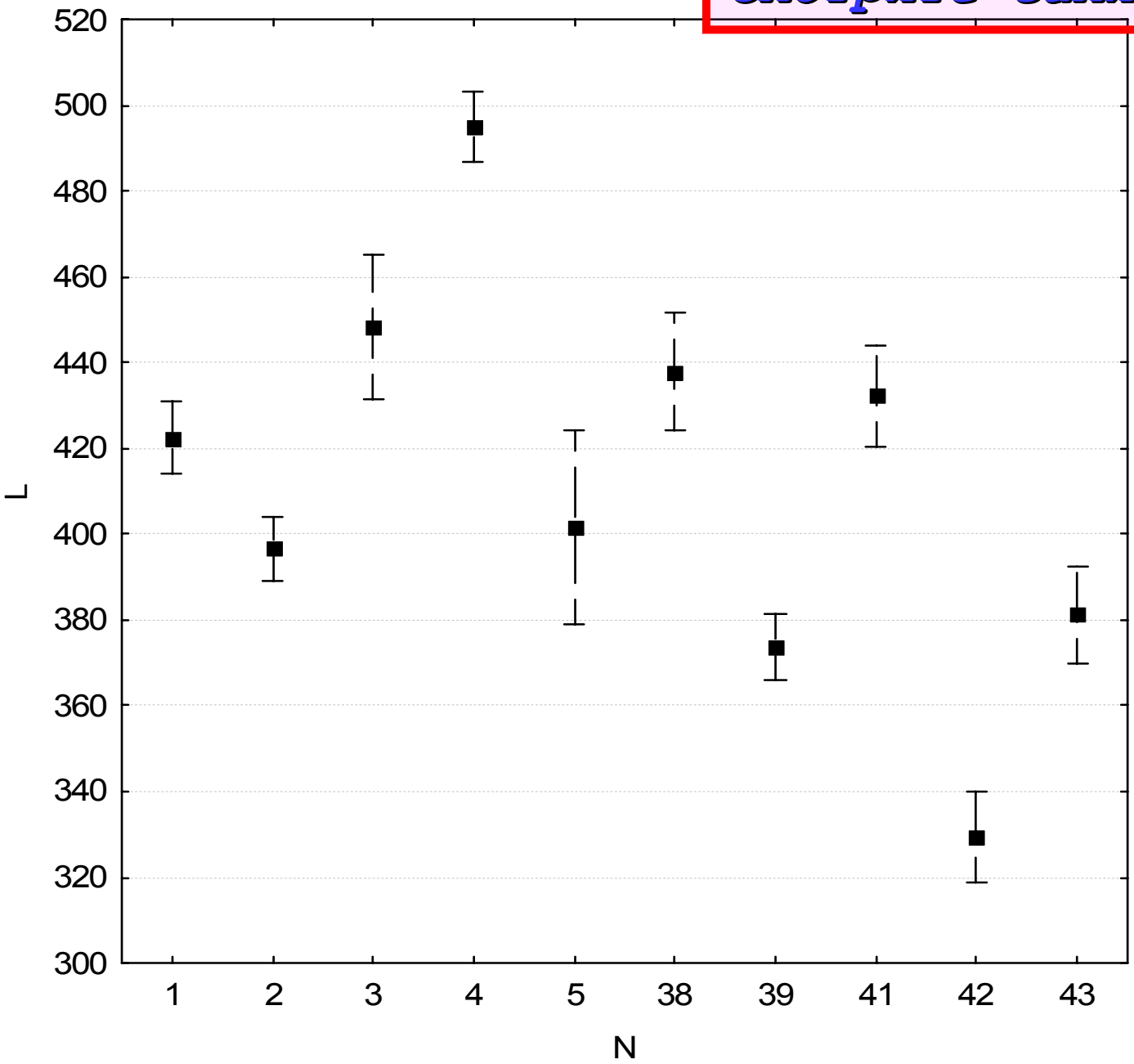
Широко применяются «точечные диаграммы» = scatter-plot

Если имеются не отдельные значения, а ряд групп (выборок, вариантов…) -


Каждая точка –
без оценки!!!.....
500
450
400
350
300
L
1
2
3
4
5
38
39
41
42
43
N

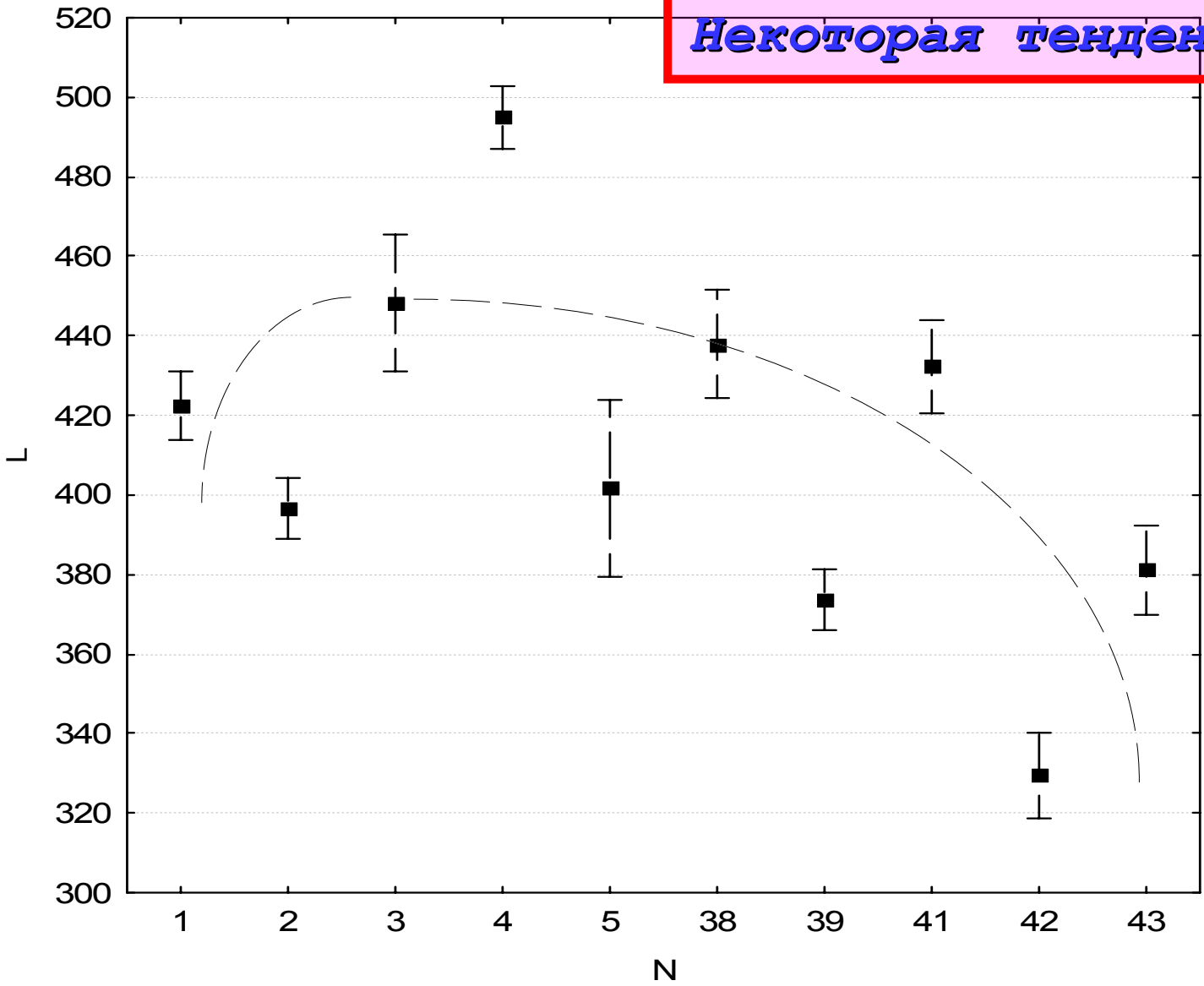

Каждая точка!!!.....
(но – доверительный интервал!)
500
450
400
350
300
L
1
2
3
4
5
38
39
41
42
43
N

Смотрите сами.....



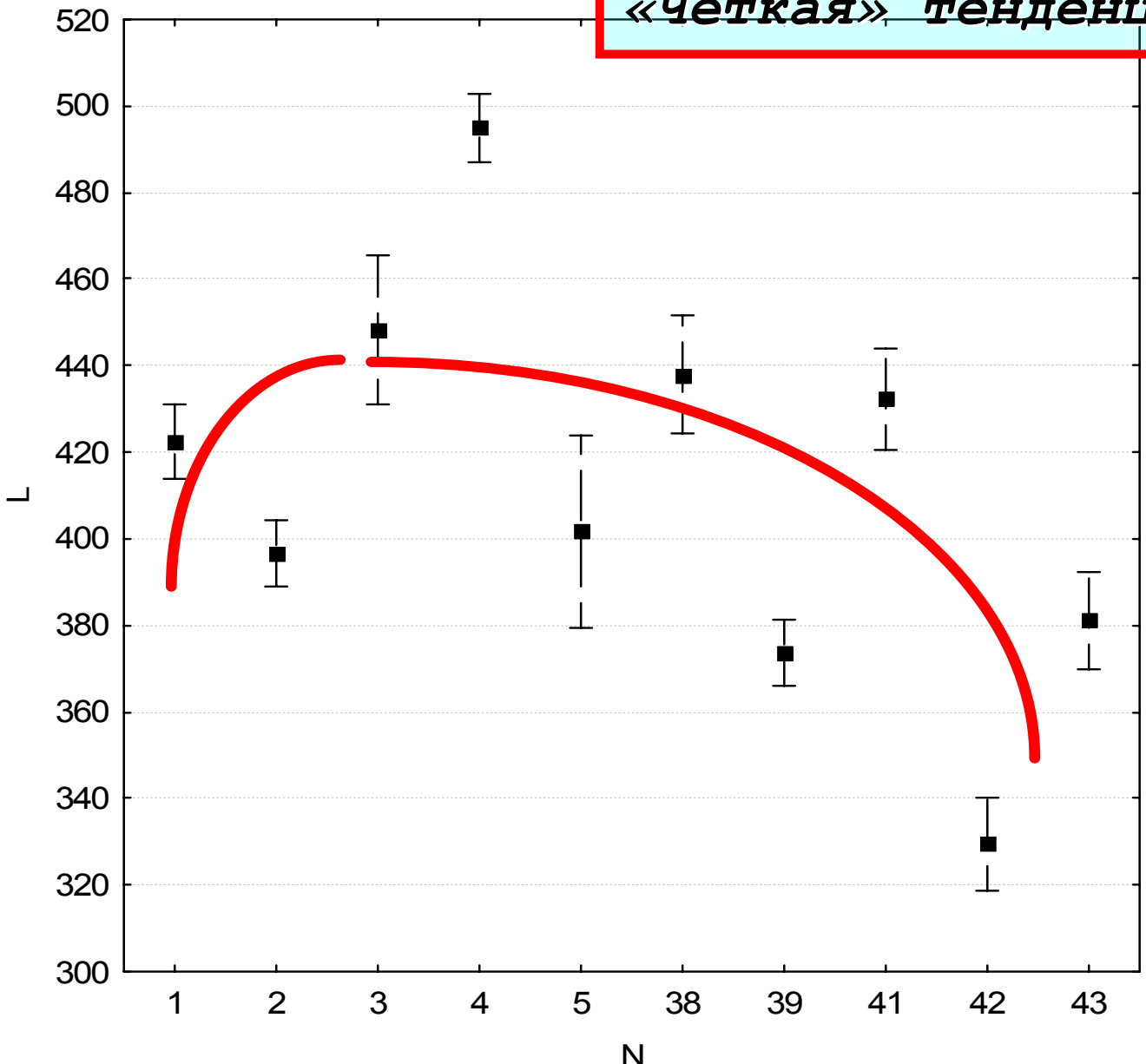
ГИПОТЕЗЫ...

Некоторая тенденция...



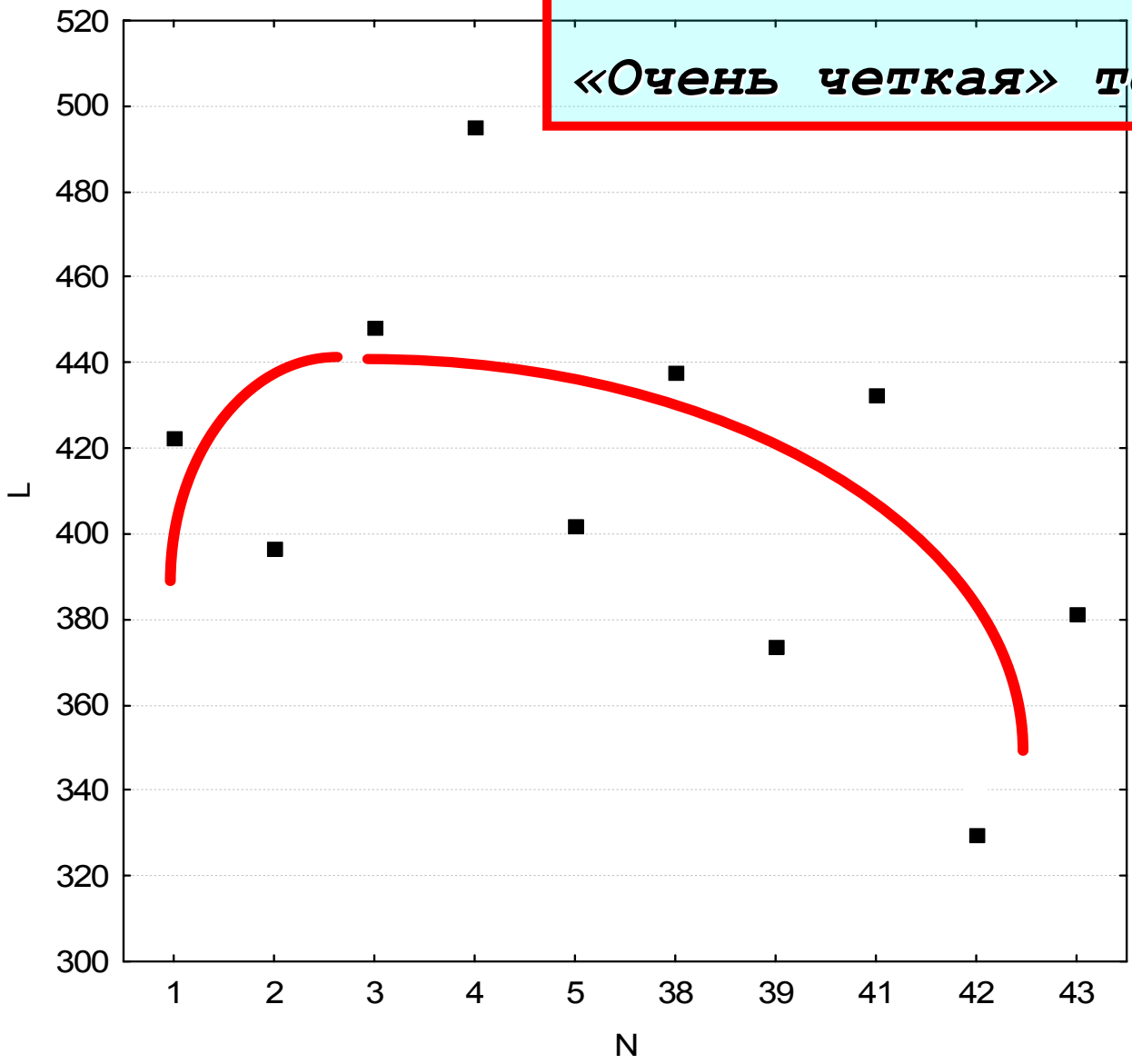
ГИПОТЕЗЫ…

«Четкая» тенденция…



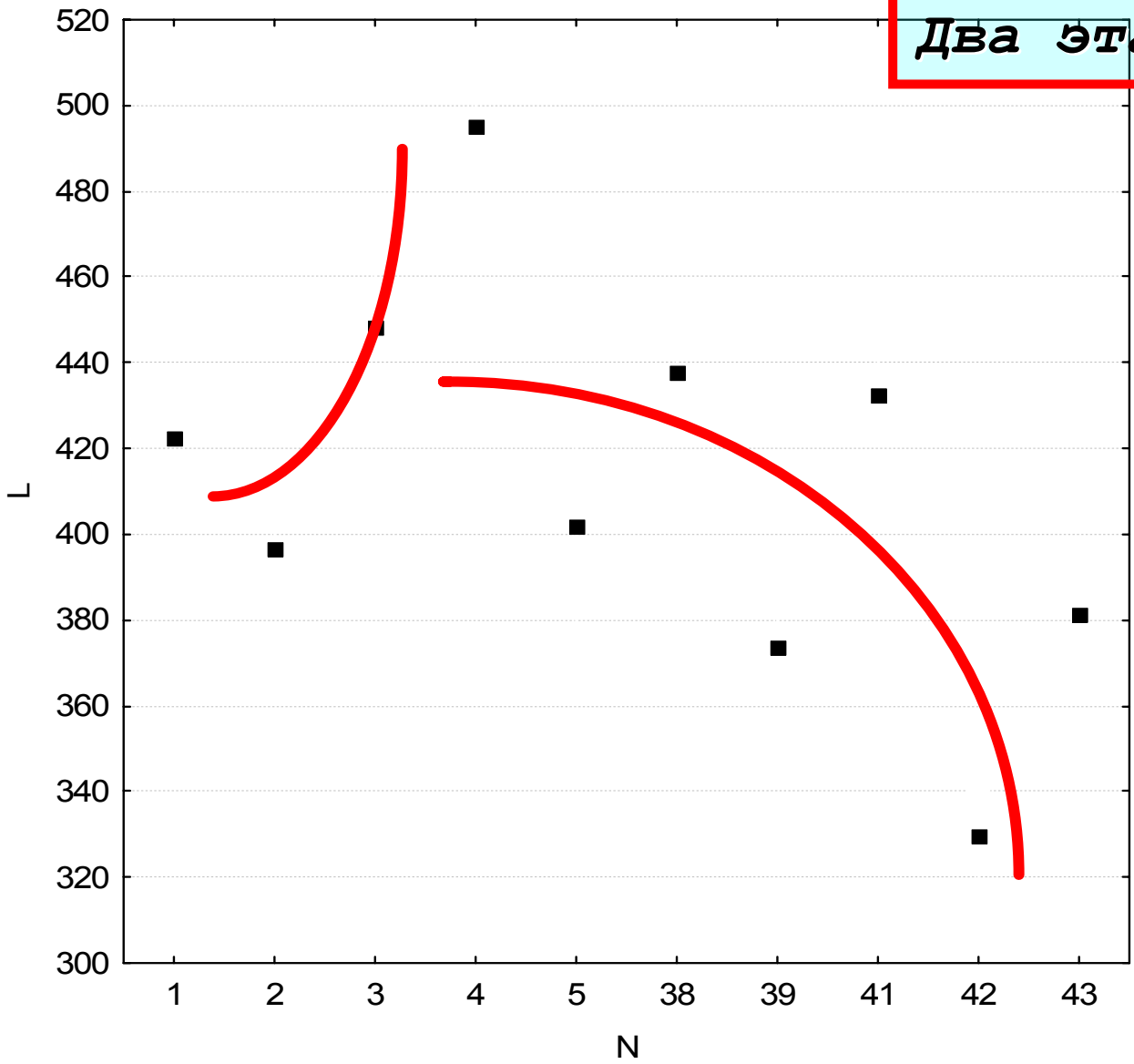
ГИПОТЕЗЫ...

«Очень четкая» тенденция...



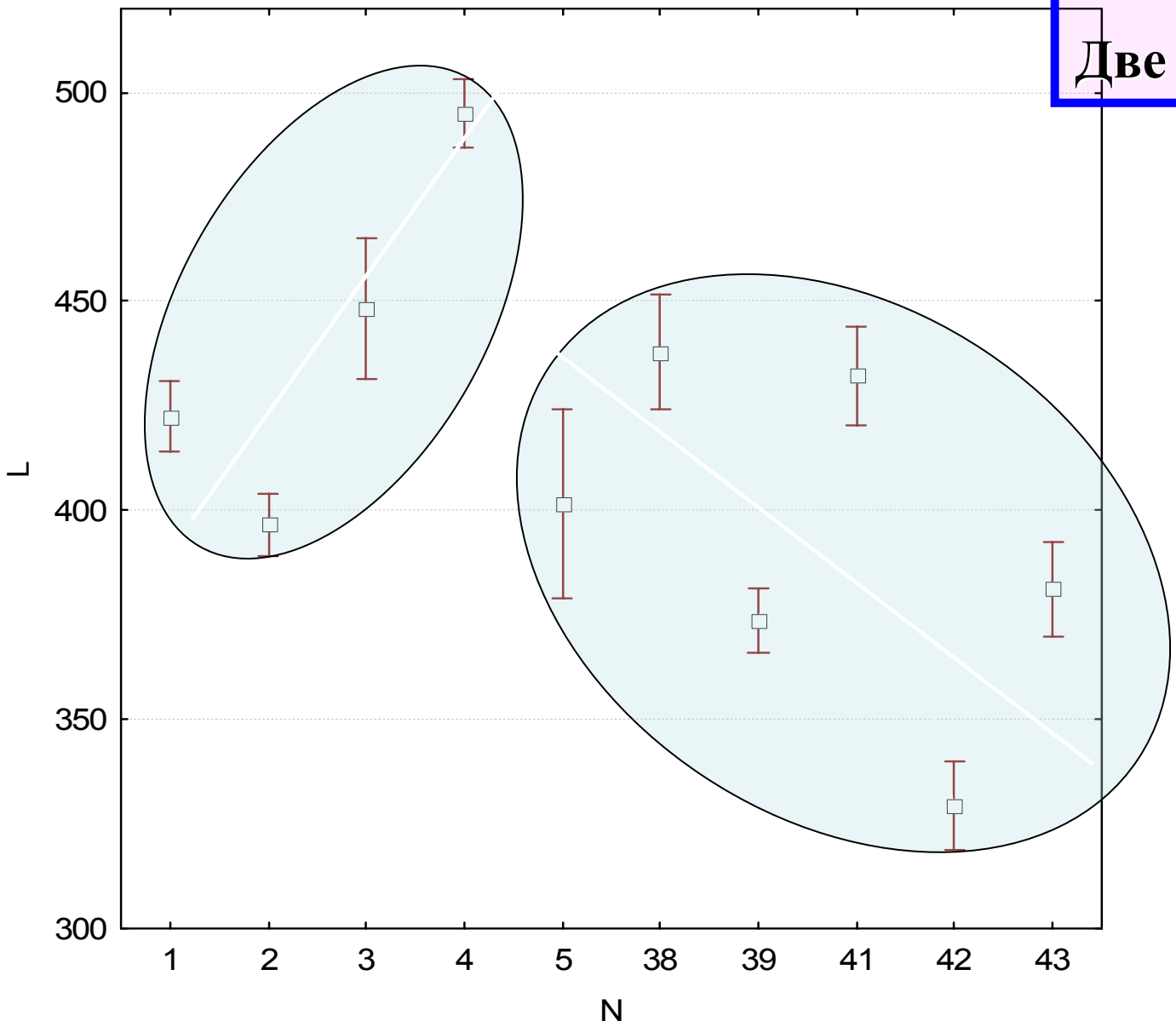
ГИПОТЕЗЫ –

Два этапа...

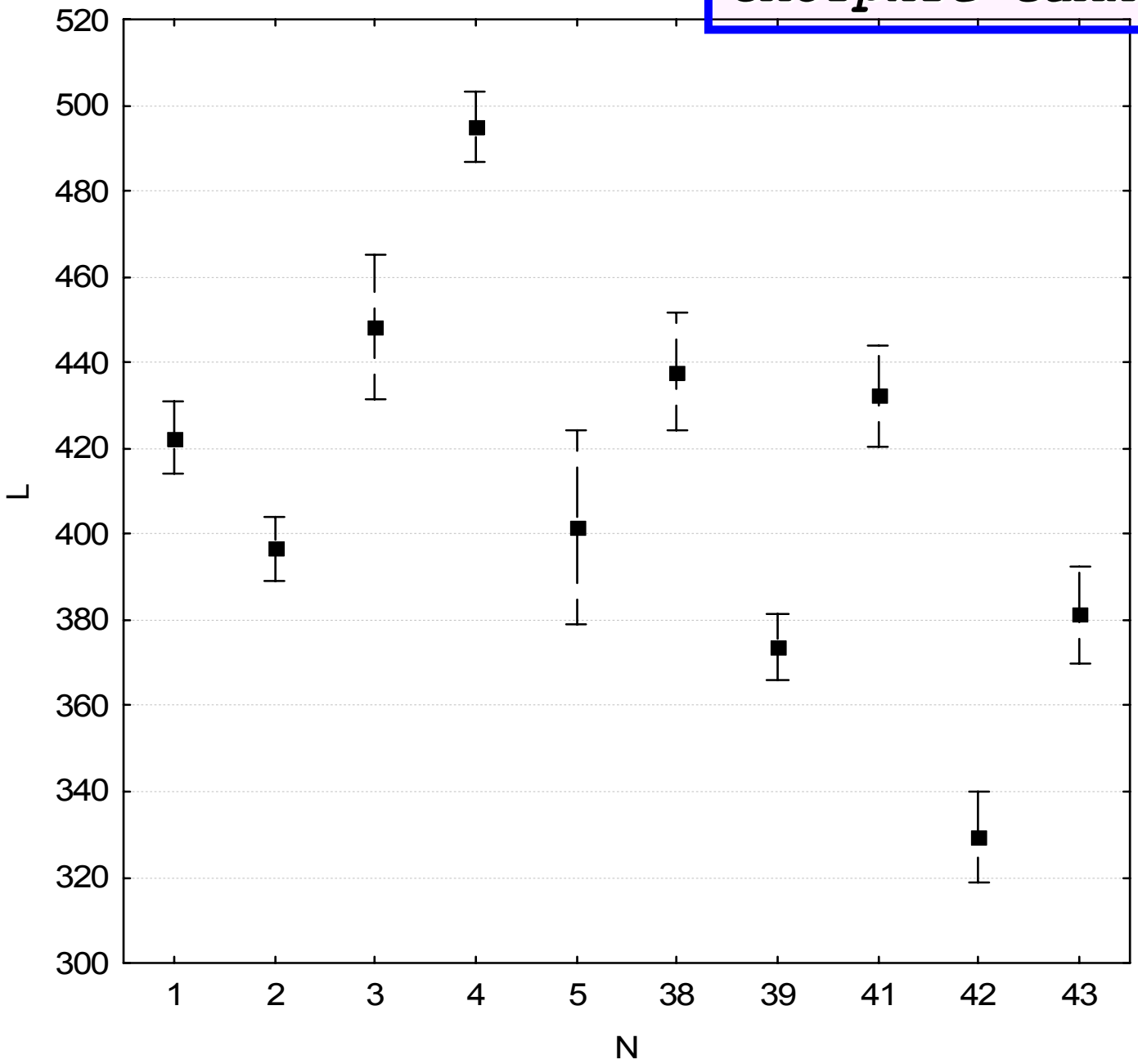


ГИПОТЕЗЫ… -

Две области….



Смотрите сами.....



- Обязательно показывать доверительные интервалы

- Не придавать значения отдельным точкам (по крайней мере - без веских оснований) – важны тенденции!

- Учитывать = показывать (и обсуждать) ВСЕ возможные гипотезы

- В дальнейшем исследовании ПРОВЕРЯТЬ и ДОКАЗЫВАТЬ

Г.Г.Винберг (1980)

Условия корректного применения в биологии элементарных эмпирических формул (Колич. методы в экологии животных, Л.,1980, с.34-36)

1. Предпочтение следует отдавать, во-первых, **формулам, приложимым ко всем или к большей части относящихся к ним материалам разных авторов**

2. Нередко результаты отдельных наблюдений за проявлениями одной и той же биологические закономерности выражают с помощью разных элементарных функций. Это ведет к накоплению несравнимых или трудно сравнимых формул, что резко снижает эффективность исследований. Необходимо **достигнуть договоренности об единообразных способах математического выражения каждой из изучаемых зависимостей**.

- 3. Практика исследований показала существенную особенность биологических данных. Материалы, полученные при, казалось бы, идентичных условиях, часто статистически достоверно различаются. Поэтому, помимо статистической обработки наблюдений, весьма важно устанавливать, **в какой мере воспроизводимы полученные количественные зависимости, к какому кругу объектов и каким условиям они приложимы**.

4. **Количественному выражению подлежат достаточно однородные по отношению к изучаемому фактору биологические материалы**. Этим важным условием плодотворности устанавливаемых количественных соотношений очень часто пренебрегают. В результате получают формально правильные, но биологически бессодержательные и ненужные, часто неоправданно усложненные математические выражения.

- 5. В практике исследований зависимости разных взаимосвязанных функций организма (или в более общей форме - разных взаимосвязанных элементов биологической системы) от некоторого фактора чаще всего изучаются раздельно. Результаты изучения каждой отдельной функции организма выражают в виде соответствующего уравнения. Сопоставление этих уравнений между собой, как показывают конкретные примеры, может приводить к абсурдным выводам. **Следовательно, зависимость от определенного фактора разных взаимосвязанных функций организма или системы надо устанавливать на одном и том же объекте и при одинаковых условиях**. При раздельном изучении функций нужно принимать во внимание необходимость согласования получаемых результатов с результатами изучения зависимости от рассматриваемого фактора **других взаимосвязанных функций** организма или элементов системы.

Линейные и нелинейные

- Внутренне линейные функции.

$$Y=\exp(\theta 1 + \theta 2\, t2 + \varepsilon)$$

- Внутренне линейные функции можно преобразовать к линейному виду.

- Например: $y = ax^2$

при замене $y - \lg(y)$

$x - \lg(x)$

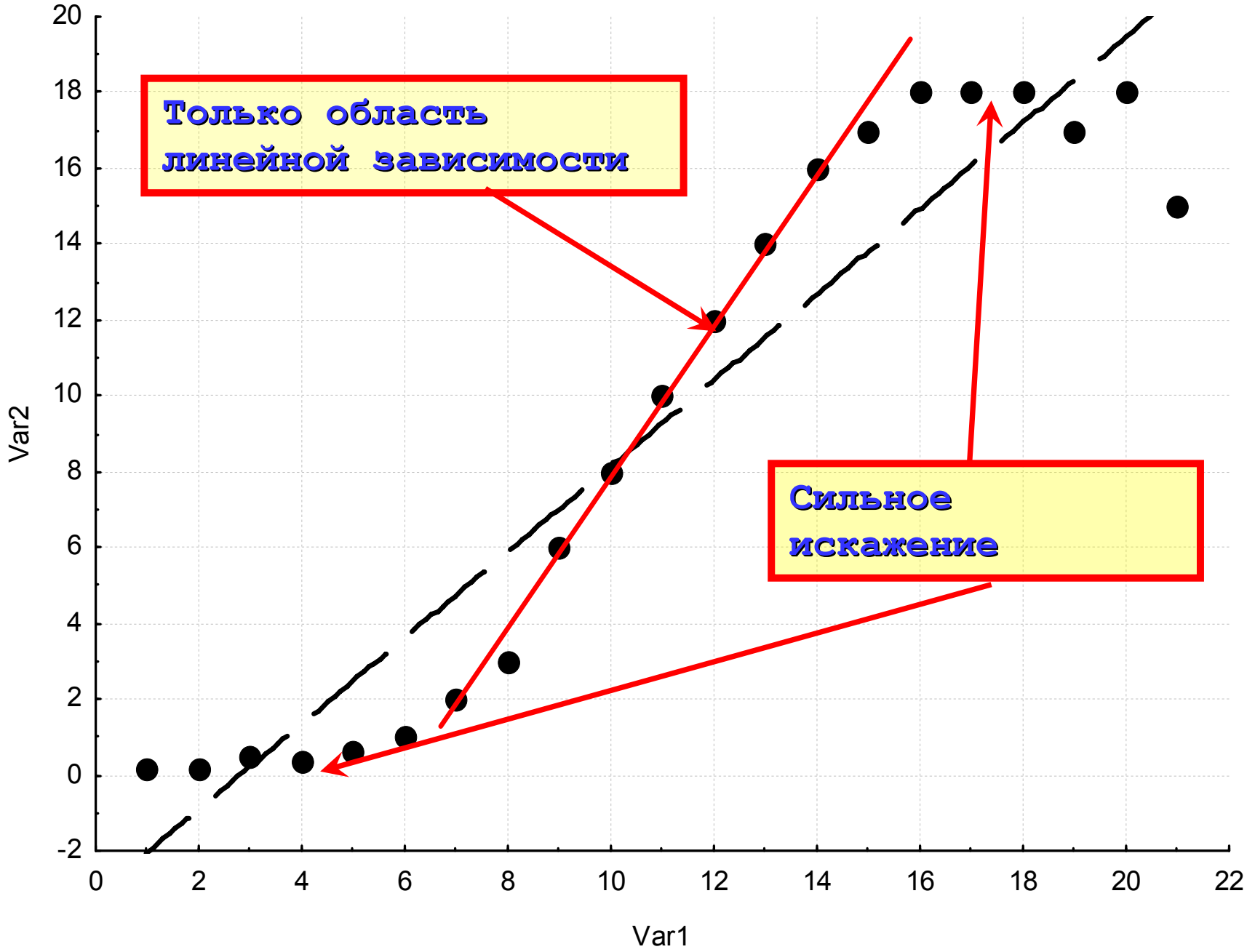
принимает линейный вид

Подбор варианта линеаризации

- **Шмидт В.М. Математические методы в ботанике. Изд. ЛГУ. 1984 (с. 101)**
- **Терентьев П.В., Ростова Н.С. Практикум по биометрии. Изд. ЛГУ. 1977 (с. 100-101)**

- «Опасность нелинейности» зачастую преувеличивается!

- Степень точности (неточности) измерений в биологических исследованиях может превышать «искажения от нелинейности»
- Обычно интервал имеющихся значений находится в области линейной зависимости


Только область линейной зависимости
Сильное искажение
Var2
Var1
20
18
16
14
12
10
8
6
4
2
0
-2
0
2
4
6
8
10
12
14
16
18
20
22

Примеры нелинейных связей

Анализ роста

- **«Соотносительный рост» – аллометрия (Huxley, 1932)**

$$y = bx^{\alpha}$$

Где α – «константа равновесия»;

При α **>1 –** положительная аллометрия

α **<1 –** отрицательная аллометрия

α **= 1 –** равномерный рост

Анализ роста

- ***S*-образные кривые роста**

  Логистическая функция:

$$Y = A /[(1 + 10^{a+bx}) + c]$$

***A* – окончательный размер, *a* и *b* – *константы* (определяют наклон, изгиб и точку перегиба), *c* – исходный размер**

Функция Гомпертца (несимметричная: растянутая верхняя ветвь)

$$Y = A /10^{10a+bx}$$

См. Шмидт В.М. 1984, с. 129-148.

Уравнение регрессии

$$Y = b_0 + b_1X$$

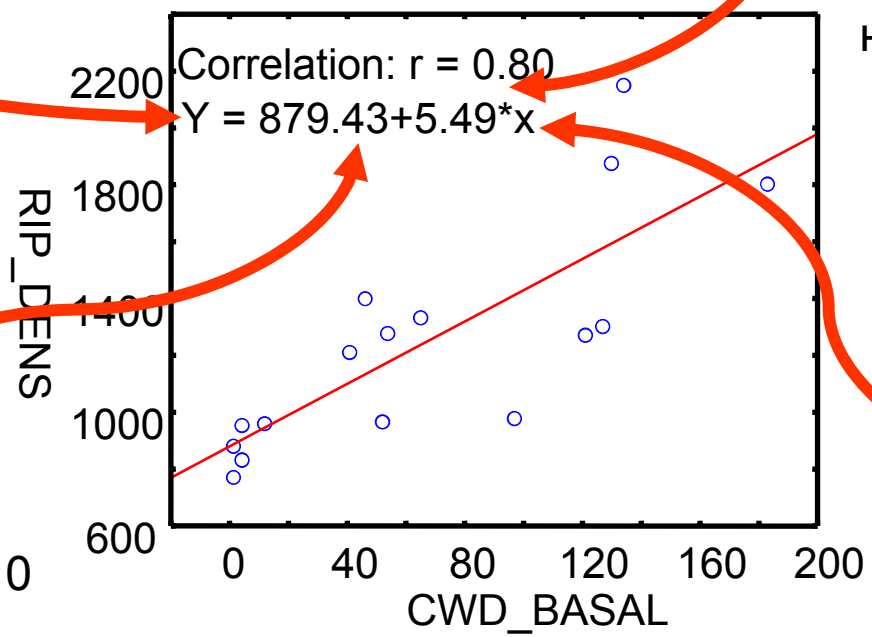
Y – зависимая переменная, отклик
Оценка μ(yi)
dependent variable, response variable

b_1 – угол наклона графика по отношению к оси X,

среднее изменение Y на единицу изменения X в выборке

Оценка β_1

slope



b_0 – ожидаемое значение Y при X = 0

Оценка β_0

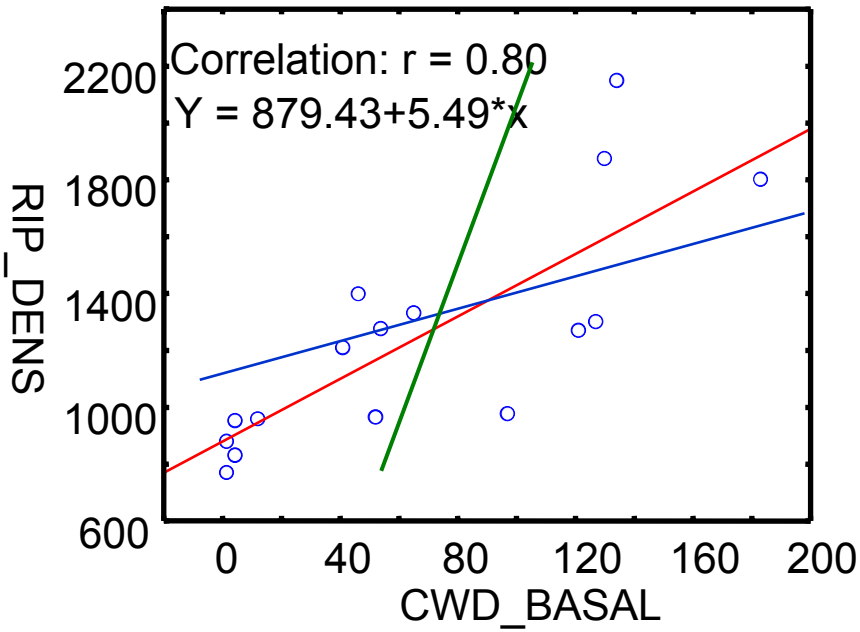
intercept

X – независимая переменная, предиктор, фактор

independent variable, predictor

Какую линию выбрать?

- На графике рассеяния можно провести множество линий, которые проходят через точки данных



- Для полученной линии регрессии

ДОВЕРИТЕЛЬНАЯ ЗОНА

и –

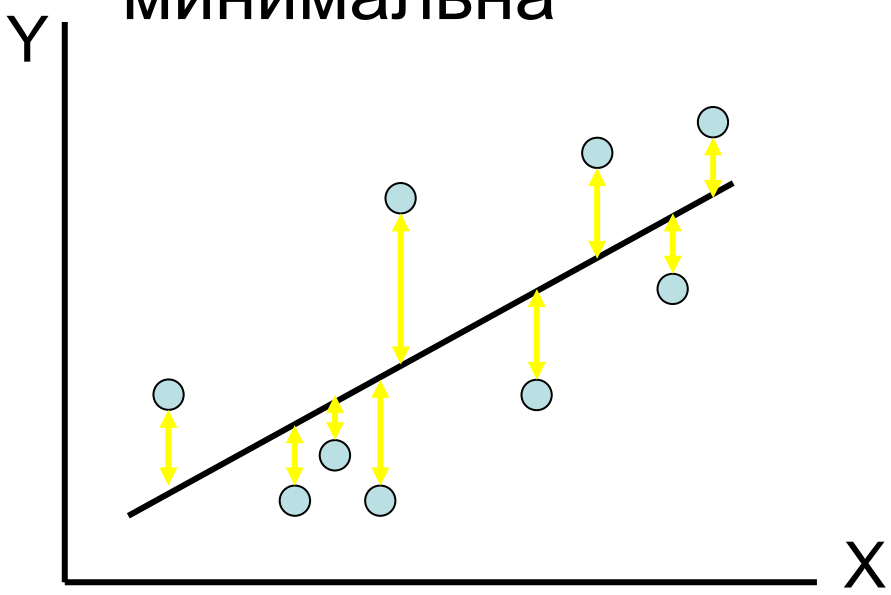
доверительные интервалы для каждого из коэффициентов уравнения

Сравнение двух линий регрессии

Урбах В.Ю. Статистический анализ в биол. И медицинских исследованиях. М. 1975. (с.203-220)

Метод наименьших квадратов

- Функция потерь
- Loss = $\sum(y_i - y_{i\,exp})^2$
- Сумма квадратов отклонений наблюдаемых от ожидаемых значений должна быть минимальна



Наблюдаемые значения Y при данном X

Предсказанные регрессией значения Y при данном X

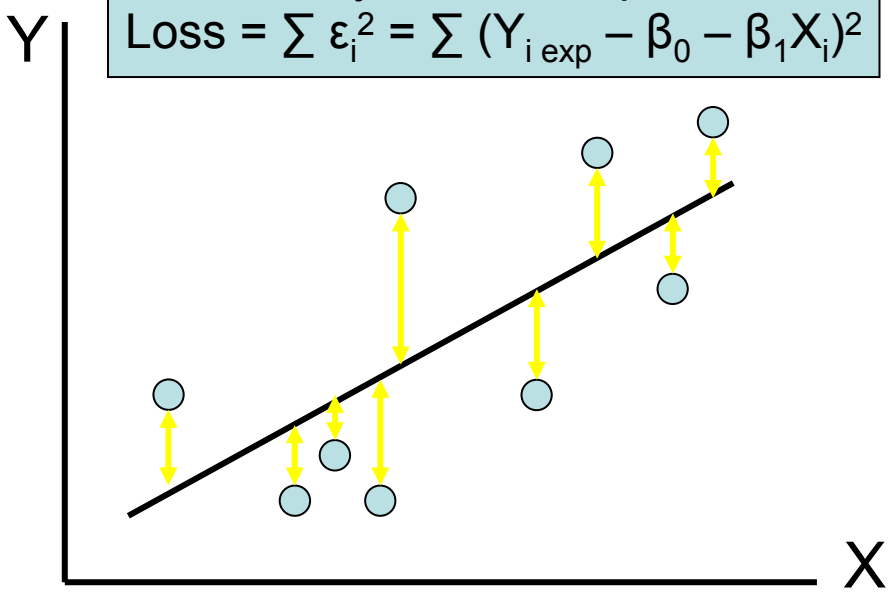
Ошибки – отклонения наблюдаемых значений от предсказанных регрессией

Рассчет коэффициентов уравнения линейной регрессии

- Модель $Y_{i\,exp} = \beta_0 + \beta_1 X + \varepsilon_i$
- Оценка модели $y_{i\,exp} = b_0 + b_1 x_i$

Функция потерь

$Loss = \sum \varepsilon_i^2 = \sum (Y_{i\,exp} - \beta_0 - \beta_1 X_i)^2$



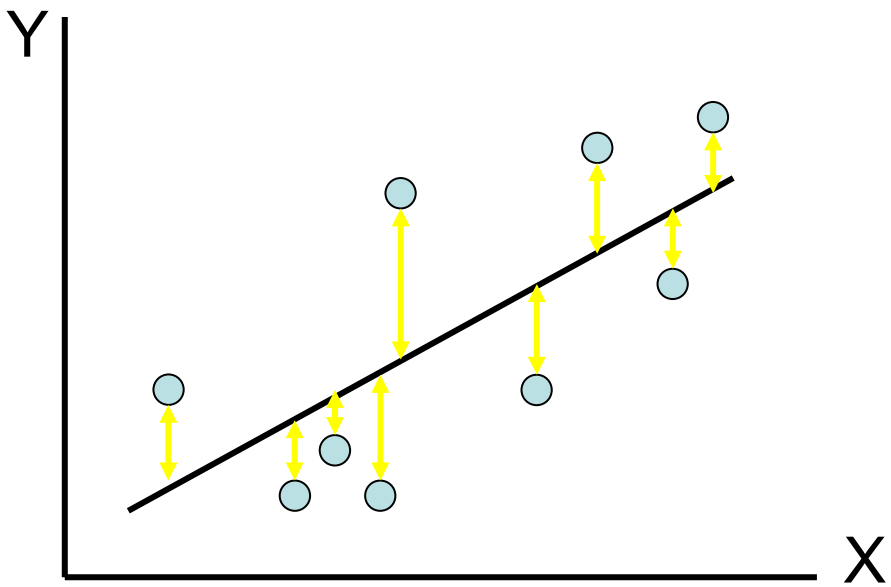
- Нужно минимизировать значение функции потерь
- Берем производные первого порядка от функции потерь по β_0 и β_1 и приравниваем их к нулю

Рассчет коэффициентов уравнения линейной регрессии

•Система т. наз. нормальных уравнений

•$-2\sum (Y_{i\,exp} - b_0 - b_1X_i) = 0$

•$-2\sum X_i (Y_{i\,exp} - b_0 - b_1X_i) = 0$



•Коэффициенты регрессии

•$b_0 = Y - b_1X$

•$b_1 = [\sum (x_i - X)(y_i - Y)] / \sum (x_i - X)^2$

Стандартные ошибки коэффициентов

•$SE_{b0} = sqrt[mSe\{1/n + X^2/\sum(x_i - X)^2\}]$

•$SE_{b1} = sqrt[mSe/\sum(x_i - X)^2]$

- Интерпретация полученного уравнения регрессии – по коэффициентам (???)

Стандартизированные коэффициенты регрессии

- Оценка коэффициентов, которая не зависит от единиц измерения X и Y
- Как получить стандартизованные коэффициенты?
  - Умножить обычный коэффициент на отношение SD_X и SD_Y
  - или
  - Подобрать уравнение регрессии по стандартизованным X и Y

$$b_1^* = b_1 * SD_X/SD_Y$$

Структура общей изменчивости

Общая изменчивость

$\sum(y_i - Y)^2$

=

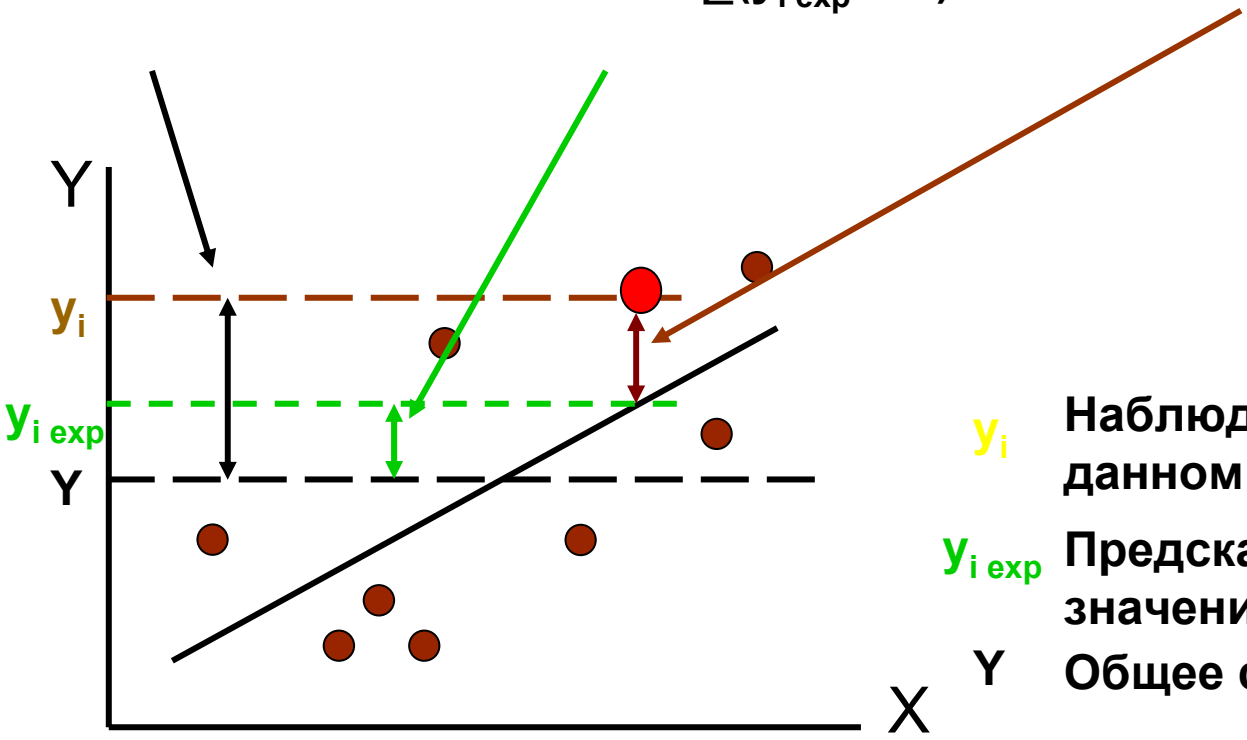
Изменчивость относительно регрессионной прямой

$\sum(y_{i\ exp} - Y)^2$

+

Остаточная изменчивость

$\sum(y_i - y_{i\ exp})^2$



y_i **Наблюдаемые значения Y при данном X**

$y_{i\ exp}$ **Предсказанные регрессией значения Y при данном X**

Y **Общее среднее значение Y**

«Особые» случаи

- *Анализ кривых «доза – эффект»* = *probit analysis* (Bliss C.)

- *«Временные ряды» = ряды динамики = Time series*

- *Анализ кривых «доза – эффект»* = probit (Bliss C.)

  в фармакологии, токсикологии… (экологии)

Литература:

1.Урбах В.Ю. Статистический анализ в биологических и медицинских исследованиях. 1975. (глава 9)

2.Беленький М.Л. Элементы количественной оценки фармакологического эффекта. 1963.

3.Зайцев Г.Н. Математический анализ биологических данных. 1991. (с.99-103)

4.Кудрин А.Н., Пономарева Г.Т. Применение математики в экспериментальной и клинической медицине. 1967.

- Варианты различаются по ДОЗЕ или ДЛИТЕЛЬНОСТИ ВОЗДЕЙСТВИЯ (количественная оценка)
- Интервалы между вариантами по интенсивности воздействия могут быть равные или неравные
- ЭФФЕКТ оценивается как число объектов в группе (варианте) с зарегистрированной реакцией (погибли – вылечились - ….)
- Группы небольшие (например, n=5-6)

- РЕЗУЛЬТАТ АНАЛИЗА --->
- LD_{50} - летальная доза для 50% выборки

  или –

  эффективная доза (ED_{50})

  эффективное время (длительность) воздействия (ET_{50}, LT_{50})

- Несколько методов, использующих
- Логарифмирование
- «Пробиты» -

  для

  а) логарифмов долей выборки, демонстрирующих наличие эффекта –

  используются

  б) накопленные частости нормального распределения
- Отсюда: probability -> probite

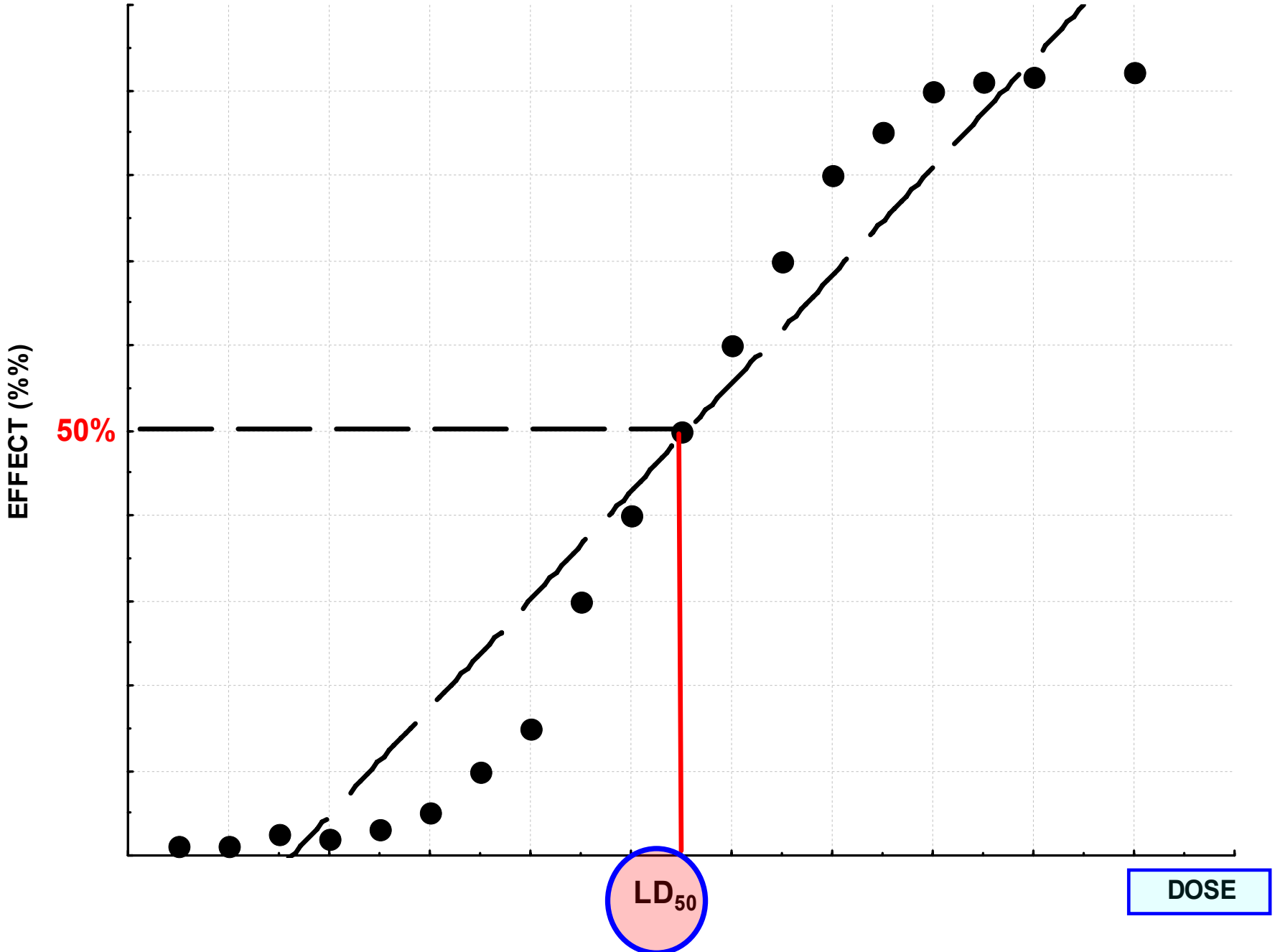
- Методы, основанные на логарифмировании, (Рида и Минча; Кербера)

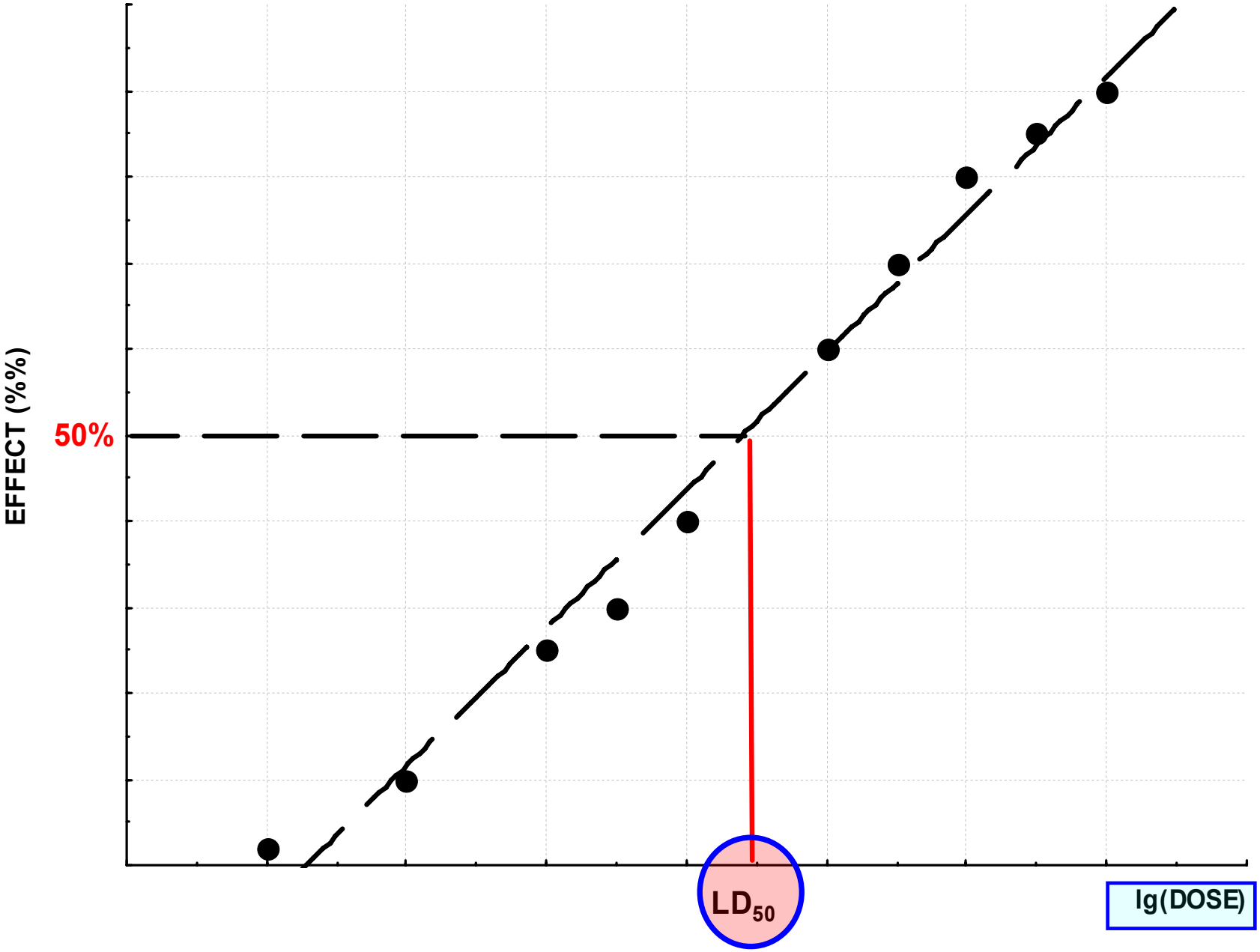
  а) более строги к данным (равноотстоящие значения доз, равенство объема групп)

  б) менее точны (недостаточная линеаризация)
- «Слабое место» пробит-анализа – допущение о нормальности кривой «доза-эффект»

Преимущества пробит-анализа

- Для величины LD_{50} имеется метод оценки ошибки и доверительного интервала (также – сравнения между этими величинами для разных воздействий)
- Основная часть вычислений может быть сделана по таблицам значений пробитов (см. пп. 1 и 3 в списке литературы)
- Показатель является *стандартным* и его можно сравнивать с результатами, полученными другми исследователями


EFFECT (%%)
50%
LD50
DOSE


EFFECT (%%)
50%
LD50
lg(DOSE)

- В пакете STATISTICA – Nonlinear Models -> Nonlinear Estimation

- Предварительно в файле данных нужно сделать логарифмирование обеих переменных

Пробит-анализ

- Доля «реагирующих» приравнивается к накопленным частостям (***z***) нормального распределения, для которых

$$Z = \Phi((x-\mu)/\sigma)$$

где Φ – интеграл вероятностей, μ и σ - математическое ожидание и стандартное отклонение распределения.

- Заменяем ***Z*** на ***p%***, ***x*** на ***lgD***, ***μ*** на $\boldsymbol{lgD_{50}}$ и получаем

$$\boldsymbol{p\% = \Phi\,((lgD - lgD_{50})/\,\sigma)}$$

- ***или*** (упрощая обозначения)–

$$\boldsymbol{p = \Phi\,((l - l_{50})/\,\sigma)} \qquad [*]$$

- После замены ***Φ*** на ***ψ*** (функция, обратная к интегралу вероятностей)

$$\boldsymbol{y' = \psi\,(p)}$$

- получаем

$$\boldsymbol{y' = (1/\sigma)l - l_{50}/\sigma}$$

- В области **p<0.5** величина ***y'*** принимает отрицательное значение. Для удобства заменяем ***y'*** на

$$y = y' + a,$$

- где ***a=5.***
- Теперь, если по оси абсцисс откладывать значения ***l*** (логарифм дозы – по вариантам), а по оси ординат $y = \psi(p) + 5$
- то точки расположатся примерно по прямой линии.
- Величина $y = \psi(p) + 5$ получила название *пробит* (от probability unit= вероятностная единица).

Для групп с объемом n 3-15 – специальные таблицы значения пробитов (не нужны не только таблицы вероятностей, но и вычисление процентов).

(часть таблицы пробитов – Урбах, 1975:245)

Число объектов в группе	Число объектов с проявляющейся реакцией						
	0	1	2	3	4	5	…15
3	3.50	4.57	5.43	6.50	–	–	
4	3.36	4.33	5.00	5.67	6.64	–	
5	3.25	4.16	4.75	5.25	5.84	6.75	
6	3.16	4.03	4.57	5.00	5.43	5.97	
7	3.10	3.93	4.43	4.82	5.18	5.57	
8	3.04	3.85	4.33	4.68	5.00	5.32	
…15	2.78	3.50	3.89	4.16	4.38	4.57	7.22

Пример

Логарифм дозы	N группы	Частота эффекта		% положит.	Накопленная частота				Пробит
		есть	нет		есть	нет	сумма	%%	
2.4	6	0	6	0.0	0	17	17	0	3.16
2.8	7	1	6	14.3	1	11	12	8.2	3.93
3.2	7	3	4	42.9	4	5	9	44.5	4.82
3.6	6	5	1	83.3	9	1	10	90.0	5.97
4.0	6	6	0	100.	15	0	15	100	6.84

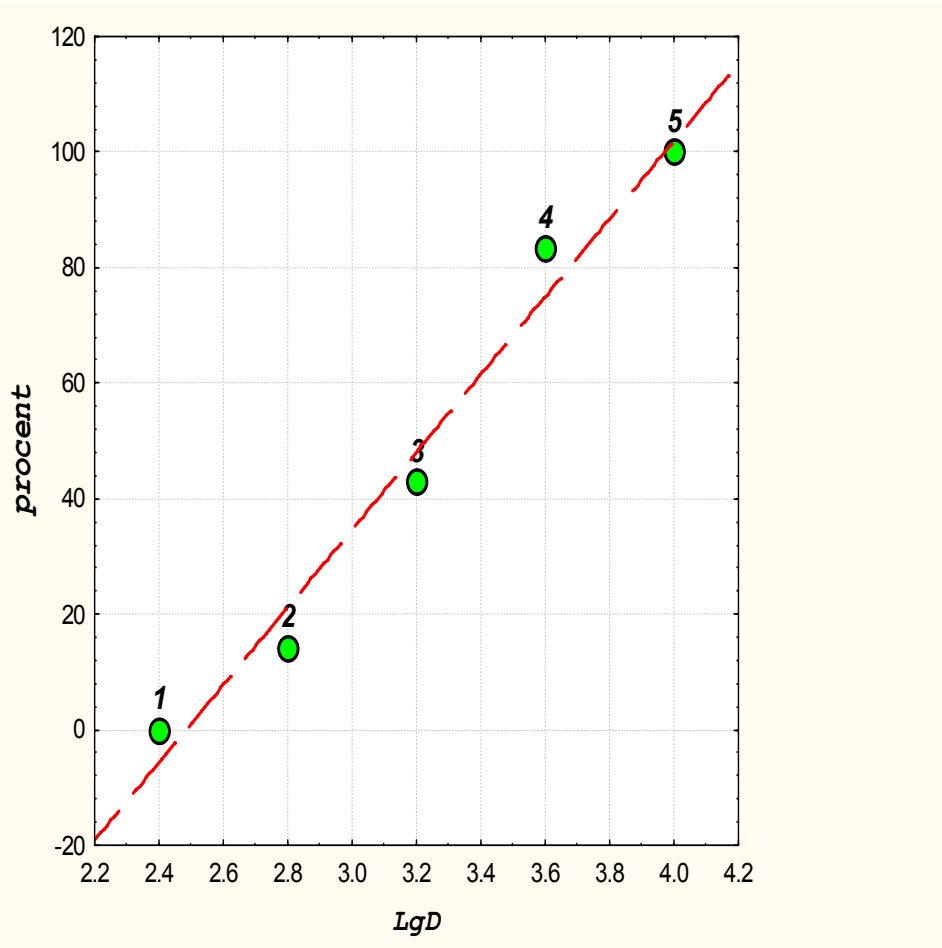
Можно просто сосчитать по «середине интервала»:

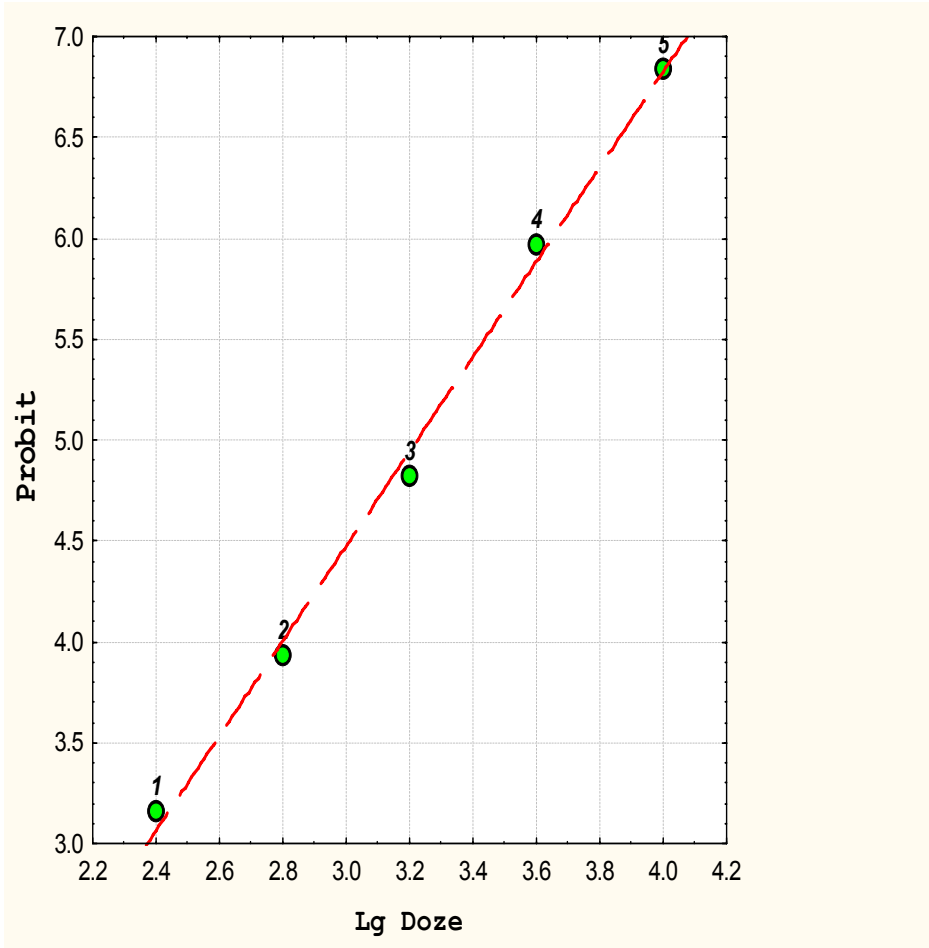
Lg $ЭД_{50}$ = 3.2 + (3.6-3.2) (50.0-42.9)/(83.3-42.9) = 3.27

Тогда $ЭД_{50} = 1.86 \cdot 10^3$.

НО – При этом мы используем только два из пяти вариантов!

А) логарифмирование
Б) пробиты





«Особые» случаи

- «Временные ряды» = ряды динамики = Time series

  а) Закономерная (фиксированная) последовательность значений в ряду значений исследуемой(ых) переменной

  корреляции между последовательными значениями в ряду (автокорреляция) и/или между рядами (кросскорреляция)

«Временные ряды» = ряды динамики = Time series

а) Фиксированная последовательность

б) Компоненты временных рядов

- Общая тенденция
- Периодическая (ие) колебания

  - их может быть несколько!

  (продолжительность общего срока и длина интервалов)
- Случайные колебания

Не только для «настоящих» рядов динамики –

- «Ряды» в пространстве
- «Ряды» метамерных органов

- **Условие:**

достаточное число членов ряда!!!

- «Сглаживание» значений во временных рядах - метод *«скользящей средней»*

  (аналогичен «линии свободной руки», но более обоснован!!!)
- Интервал сглаживания
- Коэффициенты у значений в пределах интервала (для нелинейного сглаживания)

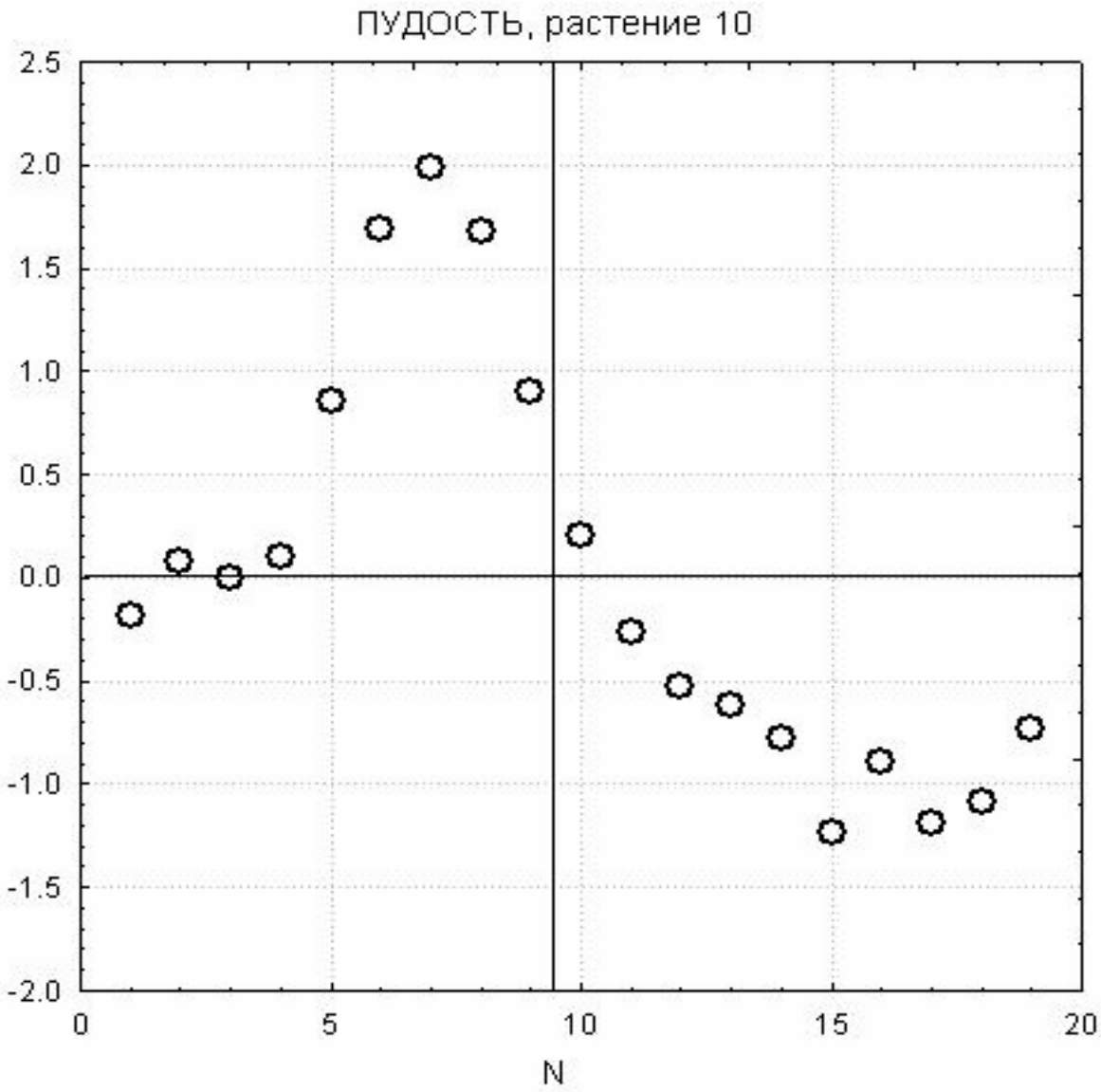
- ***Пример использования***

- Барман Ракхал Чандра «Экологическая изменчивость морфологических признаков побега *Phragmites australis* и *P.karka»*

(канд. диссертация, 1993)

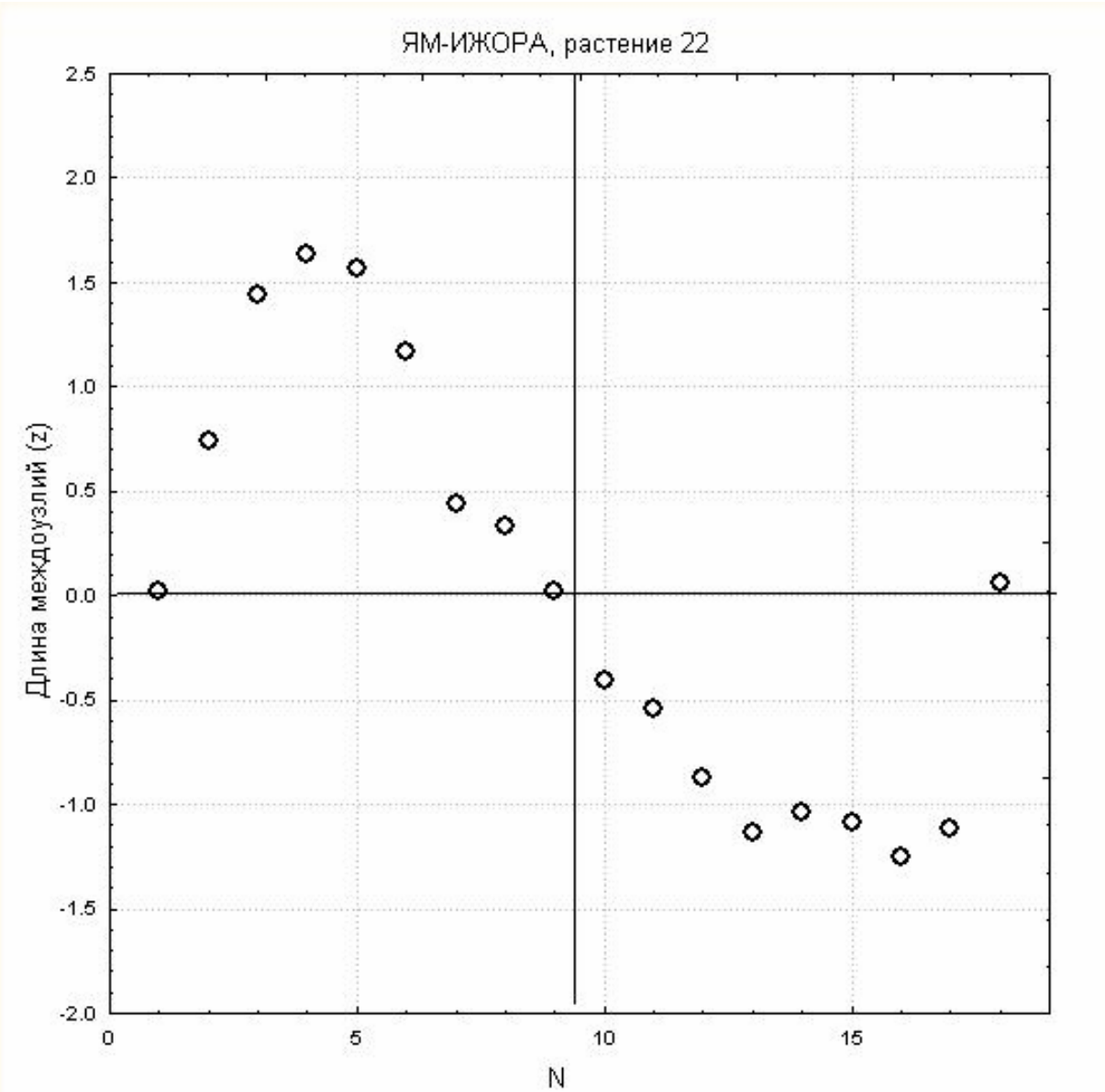
Материал: *Fragmites australis* из бассейна р. Ижора (+ Красный Бор) и из Лондона; *F. carka* из Бангладеж.

- **Измерения последовательных метамеров по побегу (N = 360)**
- **«Стандартизация» интервала (разное число метамеров!)**
- **Всего 8 признаков -> 2 компоненты (PCA)**
- **Фурье-преобразование значений компонент**
- **По коэффициентам Фурье – PCA -> ординация выборок и отдельных растений**
- **Интерпретация полученной ординации (влияние фенофазы и загрязнения)**

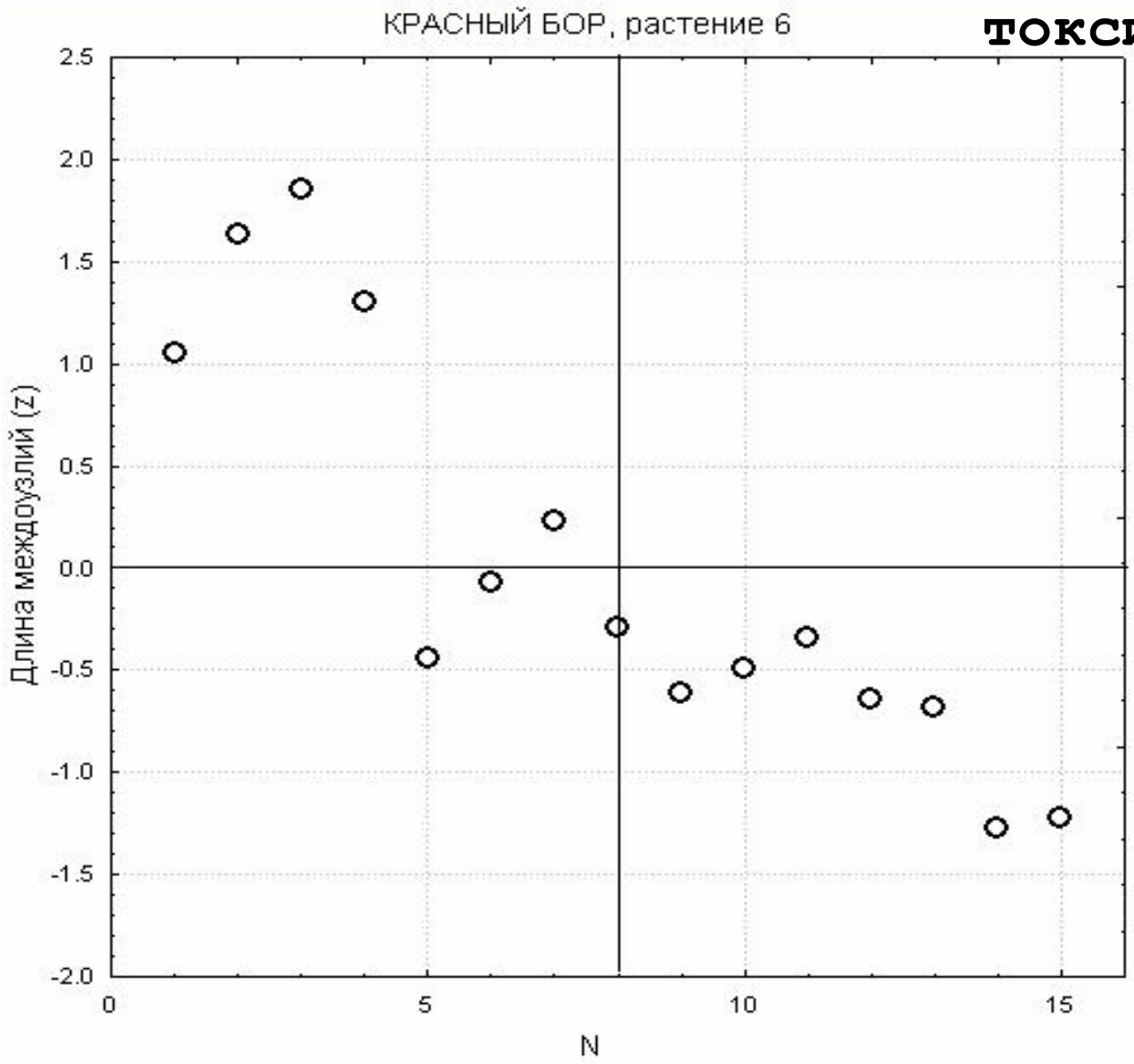


Верховья Ижоры

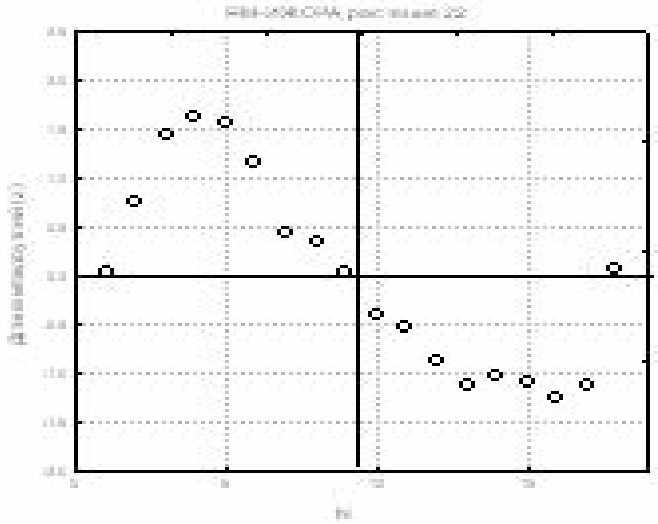
«Агрозона»


ЯМ-ИЖОРА, растение 22
Длина междоузлий (z)
N
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0
5
10
15

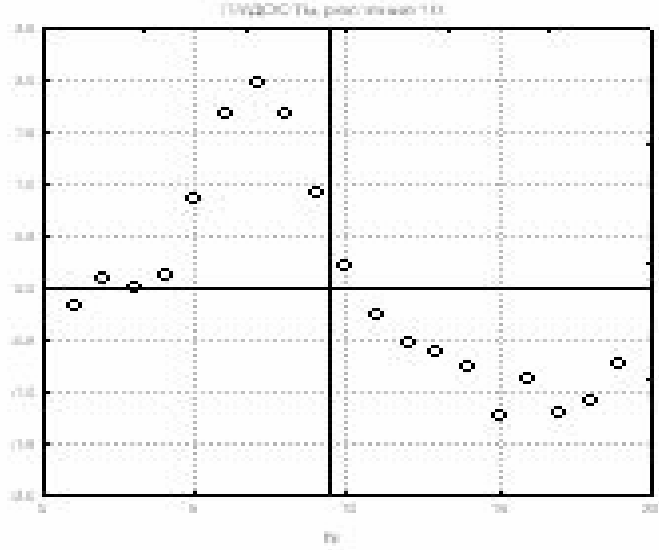
Окрестности хранилища токсичных отходов



ПРИМЕРЫ КРИВЫХ ДЛЯ РАСТЕНИЙ ТРОСТНИКА

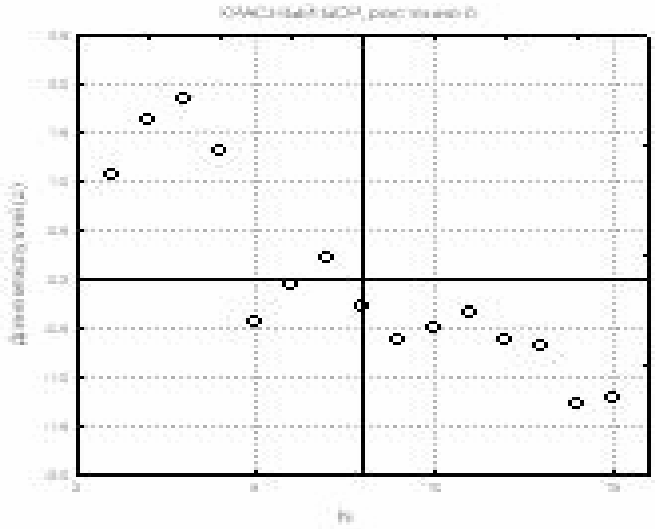


Ям-Ижора



Пудость

Красный Бор



Временные ряды = ряды динамики = time series

- Проверка последовательных значений на наличие ТРЕНДА

  Закс Л. Статистическое оценивание.1976, с. 347-356.

//Обратите внимание:

очень полезный справочник!!!//

Отношение «разбросов» (дисперсий) разностей (метод Neumann, Moore)

- **Общая дисперсия сравнивается с дисперсией последовательных разностей** (по SS)

$$SS_{total} \text{ и } SS_{(i;\, i+1)}$$

(напомним: последовательность – *фиксирована*!)

- Если последовательные значения **независимы**, то

$$SS_{(i;\, i+1)} \approx 2SS_{total} \text{ т.е.}$$

$$SS_{(i;\, i+1)} / SS_{total} \approx 2.0$$

(тренда нет ≥2.0)

- Если **есть тренд**, то

$$SS_{(i;\, i+1)} / SS_{total} < 2.0$$

(обе величины – суммы квадратов разностей: между «соседними» значениями и - со средним)

$\lvert x_i - X \rvert$	x_i	$\lvert x_i - x_{i+1} \rvert$
|2 – 8.3|= 6.3	2	|2 – 3|= 1
|3 – 8.3|= 5.3	3	|3 – 5|= 2
|5 – 8.3|= 3.3	5	|5 – 6|= 1
|6 – 8.3|= 2.3	6	|6 – 7|= 1
|7 – 8.3|= 1.3	7	|7 – 9|= 2
|9 – 8.3|=0.7	9	|9 – 10|= 1
|10 – 8.3|=1.7	10	|10 – 12|= 2
|12 – 8.3|=3.7	12	|12 – 14|= 2
|14 – 8.3|=5.7	14	|14 – 15|= 1
|15 – 8.3|=6.7	15	
Суммы квадратов разностей		
180.1	**SS**	21.0

Тренд явно ЕСТЬ:

21/180.1= 0.12

т.е. **<<2**

$\lvert x_i - X \rvert$	x_i	$\lvert x_i - x_{i+1} \rvert$
|5 – 8.3|= 3.3	5	|5 – 15|= 10
|15 – 8.3|=6.7	15	|15 – 2|= 13
|2 – 8.3|= 6.3	2	|2 – 6|= 4
|6 – 8.3|= 2.3	6	|6 – 12|= 6
|12 – 8.3|=3.7	12	|12 – 3|= 9
|3 – 8.3|= 5.3	3	|3 – 10|= 7
|10 – 8.3|=1.7	10	|10 – 9|= 1
|9 – 8.3|=0.7	9	|9 – 14|= 5
|14 – 8.3|=5.7	14	|14 – 7|= 7
|7 – 8.3|= 1.3	7	
Суммы квадратов разностей		
180.1	**SS**	526.0

Тренда НЕТ!!!

526/180.1>>2

Т.е. – чем меньше

сумма квадратов «последовательных разностей»

(между соседними значениями)

по сравнению с

суммой квадратов отклонений от среднего

ТЕМ БОЛЕЕ ВЕРОЯТНО НАЛИЧИЕ ТРЕНДА

Знаковый критерий Cox, Stuart-1955

- Весь ряд разделяется на 3 части (первая и третья – одинакового объема)
- Знаки разностей между последовательными значениями в первой-третьей частях:

  число плюсов или – минусов (S)
- Ожидаемое значение (если тренда нет) –

  S = n/6, его дисперсия – n/12,

  а SD = $(n/12)^{0.5}$
- Оцениваем отношение полученного и – ожидаемого значений

ПРИМЕР: Всего значений **n=22**, берем по **8** из первой и последней частей:

Первая треть	**4**	**7**	**3**	**5**	**6**	**7**	**8**	**9**
Вторая треть	**5**	**6**	**2**	**3**	**5**	**6**	**4**	**3**
Знаки разностей	**–**	**+**	**+**	**+**	**+**	**+**	**+**	**+**

$$z = (\left| S-n/6 \right| - 0.5)/ ((n/12)^{0.5})$$

Получаем:

$$z = (\left| 6-22/6 \right| - 0.5)/ (22/12)^{0.5} = 2.83/1.35 = 2\ .10$$

что соответствует P_0=0.0357

Установлен возрастающий тренд при $P_0 \leq 0.05$

- Для $n<30$ $z = (|S-n/6| - 0.5) / (n/12)^{0.5}$
- Для $n>30$ $z = (|S-n/6|) / (n/12)^{0.5}$
- Критические значения для одно-(1) и двухстороннего (2) критерия -

α	1	2
0.05	1.64	1.96
0.01	2.33	2.58

- Приблизительная оценка возможна – по графику последовательных значений

Многомерная регрессия

Прогнозы

- Эпидемий
- Численности «вредных» видов
- Изменений климата на Земле
- Медицинская диагностика
- Пренатальная диагностика (как особый случай)

Выбор «наилучшего» уравнения

- Все предикторы
- Последовательное включение (forward)
- Последовательное исключение (backward)
- Пошаговый – включение (stepwise=step by step forward)
- Пошаговый – исключение (stepwise=step by step backward)

Сейчас в стат. пакетах программ **–> >**

пошаговые = *stepwise*

Оценка «наилучшего» уравнения

По предикторам

- **По F-критерию (при включении и при исключении… – «добавка»)**
- **По множественному и частным коэффициентам детерминации (при включении и при исключении… – «добавка»)**
- **В пошаговых алгоритмах – и для всех ранее включенных (и исключенных)**

- **Дополнительная характеристика:**

Толерантность признака

$$T = 1 - R^2$$

Чем больше толерантность (то есть

- меньше детерминированность)

использованных для уравнения признаков, тем ниже «избыточность» полученных функций!!!