ДИСПЕРСИОННЫЙ АНАЛИЗ

…какие-то различия всегда можно обнаружить: задача анализа состоит в обосновании неслучайного (достоверного, значимого) характера этих различий.

Попарное сравнение (двух выборок по средним)

при равенстве дисперсий и объемов выборок

$$t = (X_1 - X_2) / [(SD^2_1 + SD^2_2) / n]^{1/2}$$

где X – средние арифметические, SD – стандартные отклонения,

n – объем выборок

Критерий Стьюдента оценивает отношение различий ***<u>между</u>*** выборками к ***<u>внутривыборочному</u>*** варьированию

Множественное сравнение (ряда выборок по средним)

- $F = MS_{between} / MS_{within}$

где *F* – вычисляемое значение критерия Фишера, *MS* – оценки дисперсии ***между выборками*** или – факторной (between) и ***внутри них*** – случайной (within).

Таким образом, ЛОГИКА ОЦЕНКИ наблюдаемых различий сходна в попарном и множественном сравнениях

Основная модель дисперсионного анализа может быть выражена:

$$x_{ij} = \mu + \alpha_j + \varepsilon_i$$

где

x_{ij} - i-тое значение в j-той выборке,

μ - общее среднее,

α_j - среднее “отличие” j-той выборки,

ε_i - случайное отклонение (внутри выборки).

Таким образом, исходным является предположение о том, что каждое реальное i-тое значение в j-той выборке можно разделить на компоненты, определяемые общим уровнем измеряемого признака (μ), принадлежностью к одной из групп (выборок, вариантов эксперимента) (α_j) и случайным варьированием (ε_i).

Тогда для ВСЕЙ ИЗМЕНЧИВОСТИ
(по группе выборок)

“Измерение” (оценка) общей изменчивости и ее отдельных составляющих осуществляется по величинам сумм квадратов отклонений - ***SS (Summ of Squares).*** Для однофакторного анализа такими величинами будут:

Общая сумма квадратов отклонений (“total”):

$$SS_t = \Sigma(x_{ij} - X..)^2$$

Факторная (межгрупповая) сумма квадратов отклонений (“*between*” или “*effect*”):

$$SS_x = \Sigma\Sigma(X_{.j} - X..)^2$$

Случайная (внутригрупповая или остаточная) сумма квадратов отклонений (“within” или “error”)

$$SS_e = \Sigma\Sigma(x_i - X_{.j})^2$$

Компоненты изменчивости

Общая изменчивость =

= Факторная + Случайная

$$SSt = SSx + SSe$$

Синонимы:

Факторная = межгрупповая (Sum of squares between/among/explained; among groups variation)

Случайная = внутригрупповая (Sum of squares within/error/unexplained; within-group variation, residual)

- Вспомним, как мы определяли (и вычисляли) величину дисперсии для одной выборки:

$$SD^2 = \Sigma (x_i - X)^2 / (n-1)$$

или

$$SD^2 = SS/df$$

где *SS* – сумма квадратов (отклонений)

Вычислительная формула ***SS***:

$$SS = \Sigma x_i^2 - (\Sigma x_i)^2/n$$

*Соответственно, вычисления **SS** в дисперсионном анализе:*

$$SSt = \sum\sum (x_{ij}^2) - (\sum x_{ij})^2/N$$

- $SSx = 1/n_j \sum((\sum x_i)^2) - (\sum x_{ij})^2/N$
- $SSe = \sum(x_{ij}^2) - 1/n_i \cdot \sum(\sum x_i)^2$

Где ***N*** - общая численность всех данных,

n – объем отдельной выборки, ***a*** – число групп.

Величина $(\sum x_{ij})^2/N$ – **«поправка» = *T*,**

$1/n_j \sum((\sum x_i)^2)$ – ***«факторная сумма» = A;***

$$\sum(x_{ij}^2) = Y$$

Тогда можно упрощенно записать:

$$SSt = Y - T$$

$$SSx = A - T$$

$$SSe = Y - A$$

Для более сложных схем:

- В двухфакторном

$$SS_x = SS_A + SS_B + SS_{AB}$$

- В трехфакторном

$$SS_x = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC}$$

т.е. усложняется вычисление факторной суммы (прибавляются выделение новых факторов и их взаимодействий)

Кроме SS нам нужно:

- ***MS*** – средние квадраты (отклонений):

$$MS = SS/df$$

Для каждой компоненты изменчивости это - оценка соответствующей дисперсии.

Определение df (числа степеней свободы)

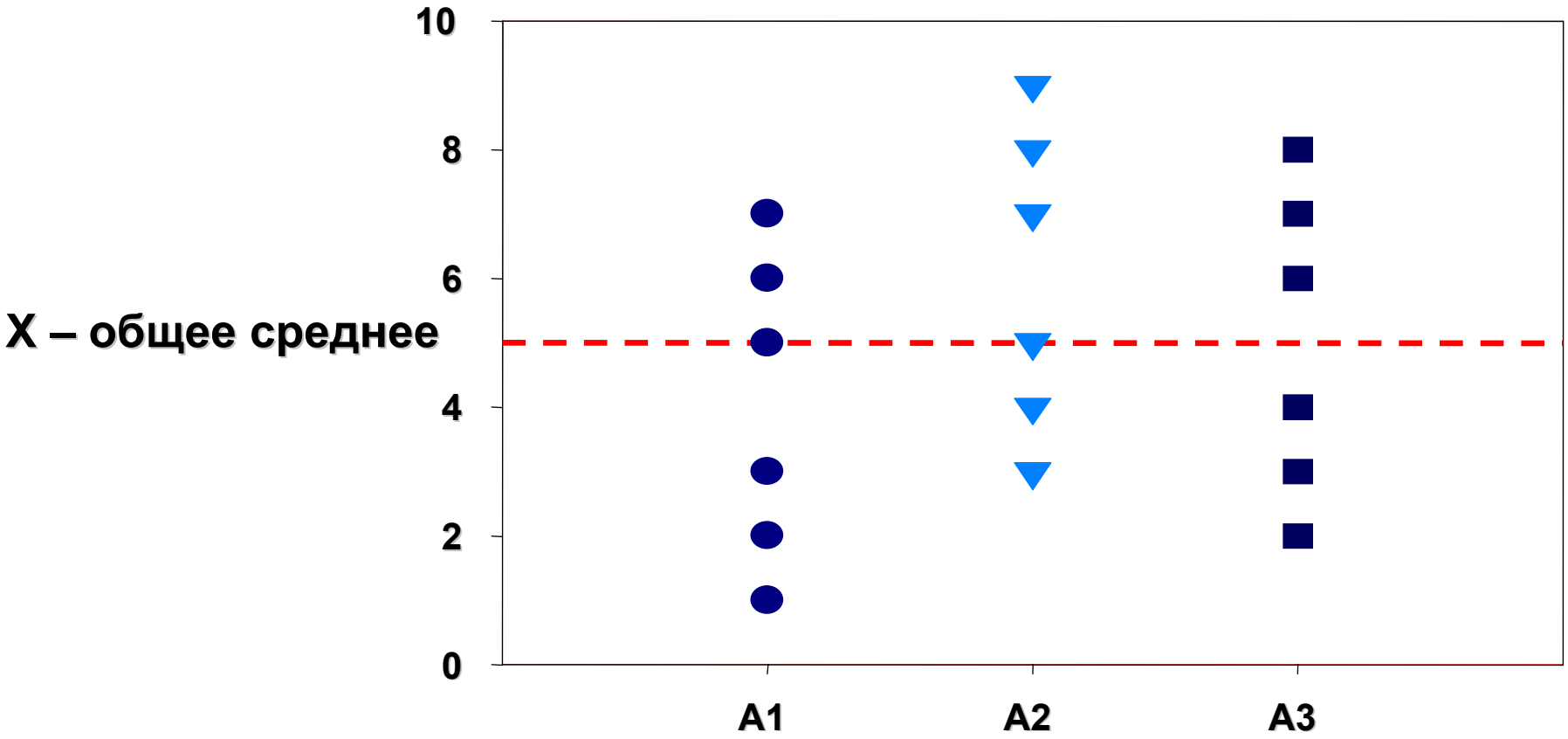
$$df_{tot} = N - 1$$

$$df_x = a - 1$$

$$df_e = a\,(n - 1) = N - a$$

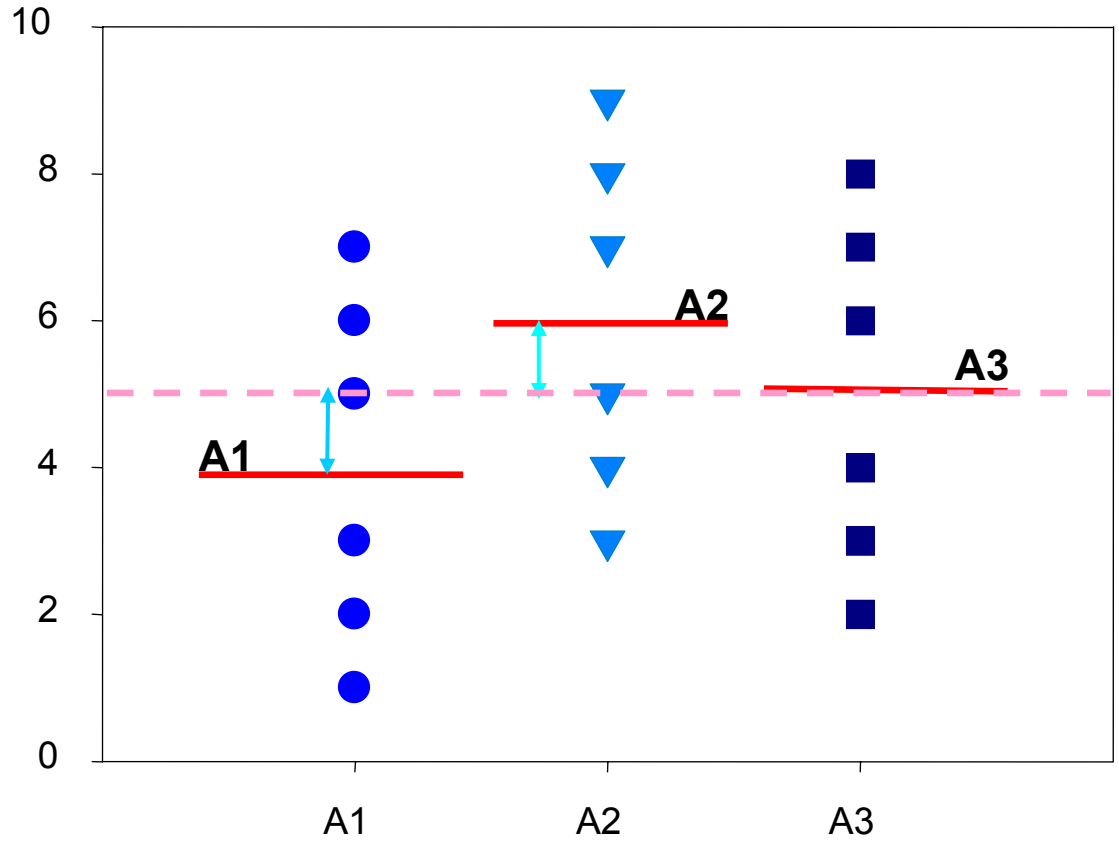
Общая сумма квадратов

$SSt = SSx + SSe$



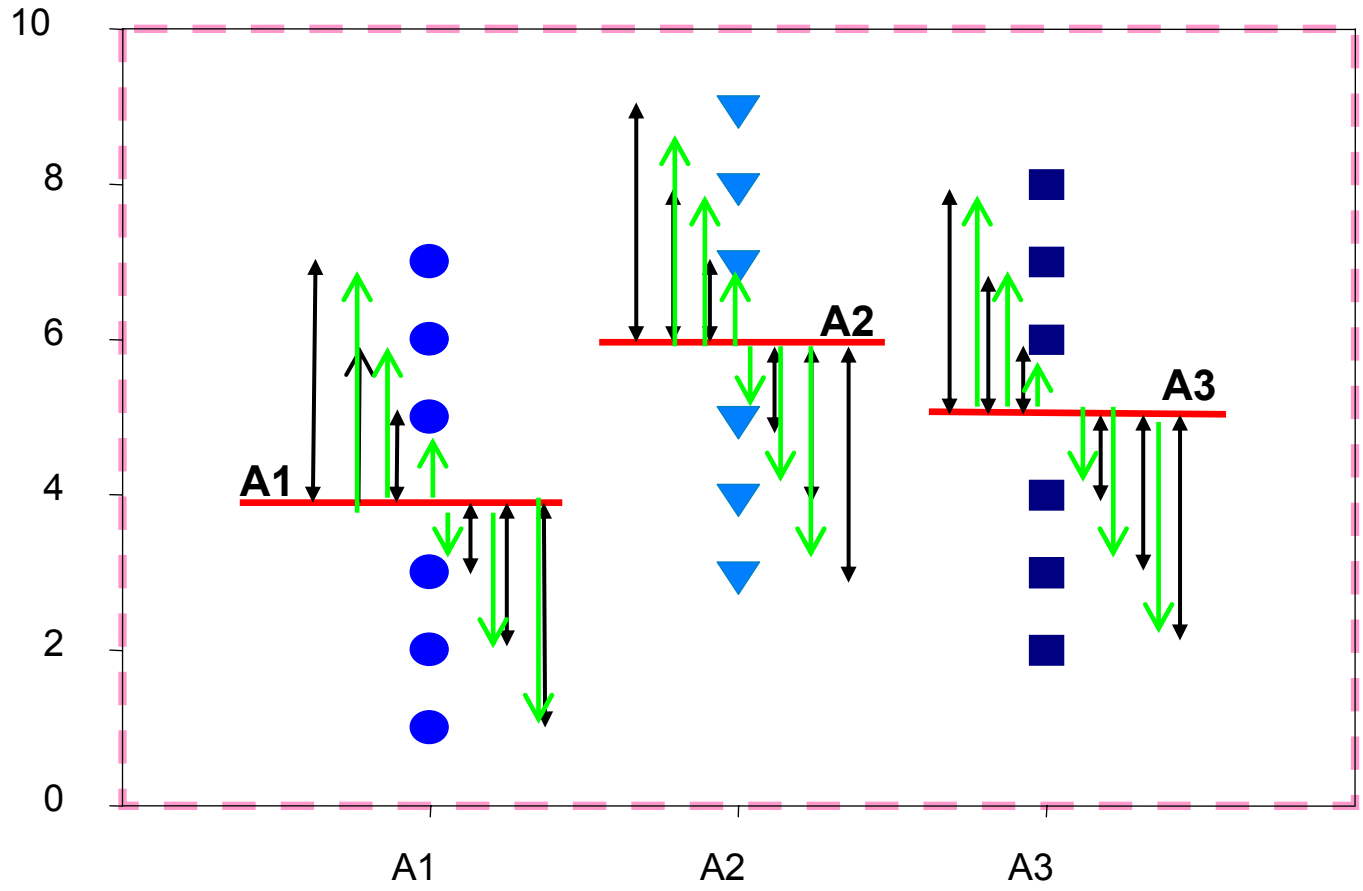
Факторная сумма квадратов

SSx



Случайная сумма квадратов

SSe



Однофакторный дисперсионный анализ: исходные данные

- **Фактор – groupA (число ошибок в работах по математике, физике, литературе)**
- **Всего групп – 3 (A1, A2, A3)**
- **В группе 6 наблюдений**
- **Всего 18 наблюдений**

groupA	numbers
A1	1 2 3 5 6 7
A2	2 3 4 8 6 7
A3	7 8 9 5 4 3

Порядок вычислений

- $\sum x_i$ (A1)=24, (A2)=36 и (A3)= 30;
- общая сумма $\sum x_{ij}$=90;
- сумма квадратов всех значений $\sum \mathbf{x_{ij}^2 = 546}$

$$SS_{total} = 546 - 90^2/18 = 546 - 450 = 96$$

$$SS_x = SS_A = 1/6 \cdot (24^2 + 30^2 + 36^2) - 90^2/18 =$$

$$= 462 - 450 = 12$$

$$SS_e = 546 - 1/6 \cdot (24^2 + 30^2 + 36^2) =$$

$$= 546 - 462 = 84$$

Таблица результатов

Изменчивость		SS	df	MS	F
x (A)	факторная	12	3 - 1=2	6.0	1.071
error (e)	случайная	84	18 - 3 =15	5.6	
Total	общая	96			

F(0.05) = 3.68; F(0.01)=6.36

Какой вывод?

Внутри групп – высокая изменчивость!!!

На самом деле можно разделить их еще и по другому фактору (B = пол):

Градации факторов	B1 мальчики	B2 девочки	Суммы A	Сумма квадратов
A1 математика	1 2 3	5 6 7	24	
	6	18		
A2 физика	2 4 3	8 6 7	30	
	9	21		
A3 литература	7 8 9	3 4 5	36	
	24	12		
Суммы B	39	51	90	546

Формулы вычислений для двухфакторного анализа

$$SSt = Y - T$$

$$SSx = AB - T$$

$$SS_A = A - T$$

$$SS_B = B - T$$

$$SS_{AB} = SSx - (SS_A + SS_B)$$

$$SSe = Y - AB$$

- **Результаты по измененным данным (с добавлением разделения по второму фактору - B):**

- $SS_x = 1/3 \cdot(6^2+24^2+9^2+18^2+12^2+21^2) - 90^2/18 =$
  $= 1602/3 - 450 = 534 - 450 = 84$
- $SS_A = 1/6 \cdot(24^2+30^2+36^2) - 90^2/18 = 462 - 450 = 12$
- $SS_B = 1/9 \cdot(39^2+51^2) - 90^2/18 = 458 - 450 = 8$
- $SS_{AB} = 84 - (12+8) = 64$
- $SS_e = 546 - 1/3 \cdot(6^2+24^2+9^2+18^2+12^2+21^2) =$
  $= 546 - 534 = 12$

Результаты двухфакторного дисперсионного анализа

	SS	df		MS	F	P_0	$\eta^2\%$	$\eta^2\%^*$
x	84	ab-1	5	16.8	16.8	0.000047	87.5	82.3
A	12	a-1	2	6.0	6.0	0.015625	12.5	11.8
B	8	b-1	1	8	8.0	0.015220	8.3	7.8
AB	64	(a-1)* (b-1)	2	32	32.0	0.000015	66.7	62.7
e	12	N-ab	12	1			12.5	17.7
To-tal	96						100	100

Вычисления «по статистикам»

$$SS_x = n_{ab} \cdot \Sigma X_{AB}^2 - \{n_{ab} \cdot (\Sigma X)^2\}/ab$$

$$SS_A = n_a \cdot \Sigma X_A^2/b - \{n_{ab} \cdot (\Sigma X)^2\}/ab$$

$$SS_B = n_b \cdot \Sigma X_B^2/a - \{n_{ab} \cdot (\Sigma X)^2\}/ab$$

$$SS_{AB} = SS_x - (SS_A + SS_B)$$

$$SS_e = (n_{ab} - 1) \cdot \Sigma SD^2$$

$$SS_t = SS_x + SS_e$$

В нашем примере:

	X		ΣX_A	SD	
	B1	B2		B1	B2
A1	2	6	8	1	1
A2	3	7	10	1	1
A3	8	4	12	1	1
ΣX_B	13	17	30		

- **Предварительные вычисления:**
- $\sum(X_{AB})^2 = 2^2 + 3^2 + 8^2 + 6^2 + 7^2 + 4^2 = 178$
- $\sum(X_A)^2 = 8^2 + 10^2 + 12^2 = 308$
- $\sum(X_B)^2 = 13^2 + 17^2 = 458$
- **Основные вычисления**
- $SSx = 3 \cdot 178 - 3 \cdot (30^2)/3 \cdot 2 = 534 - 450 = 84$
- $SS_A = 3 \cdot 308/2 - 3 \cdot (30^2)/3 \cdot 2 = 462 - 450 = 12$
- $SS_B = 3 \cdot 458/3 - 3 \cdot (30^2)/3 \cdot 2 = 458 - 450 = 8$
- $SS_{AB} = 84 - (12+8) = 64$
- $SSe = (3 - 1) \cdot (1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2) = 12$
- $SSt = 84 + 12 = 96$

$\eta^2\%$ - так называемый "коэффициент внутриклассовой корреляции" (intraclass correlation) по Фишеру. Она может быть вычислена как доля (или процент) от общей изменчивости, например:

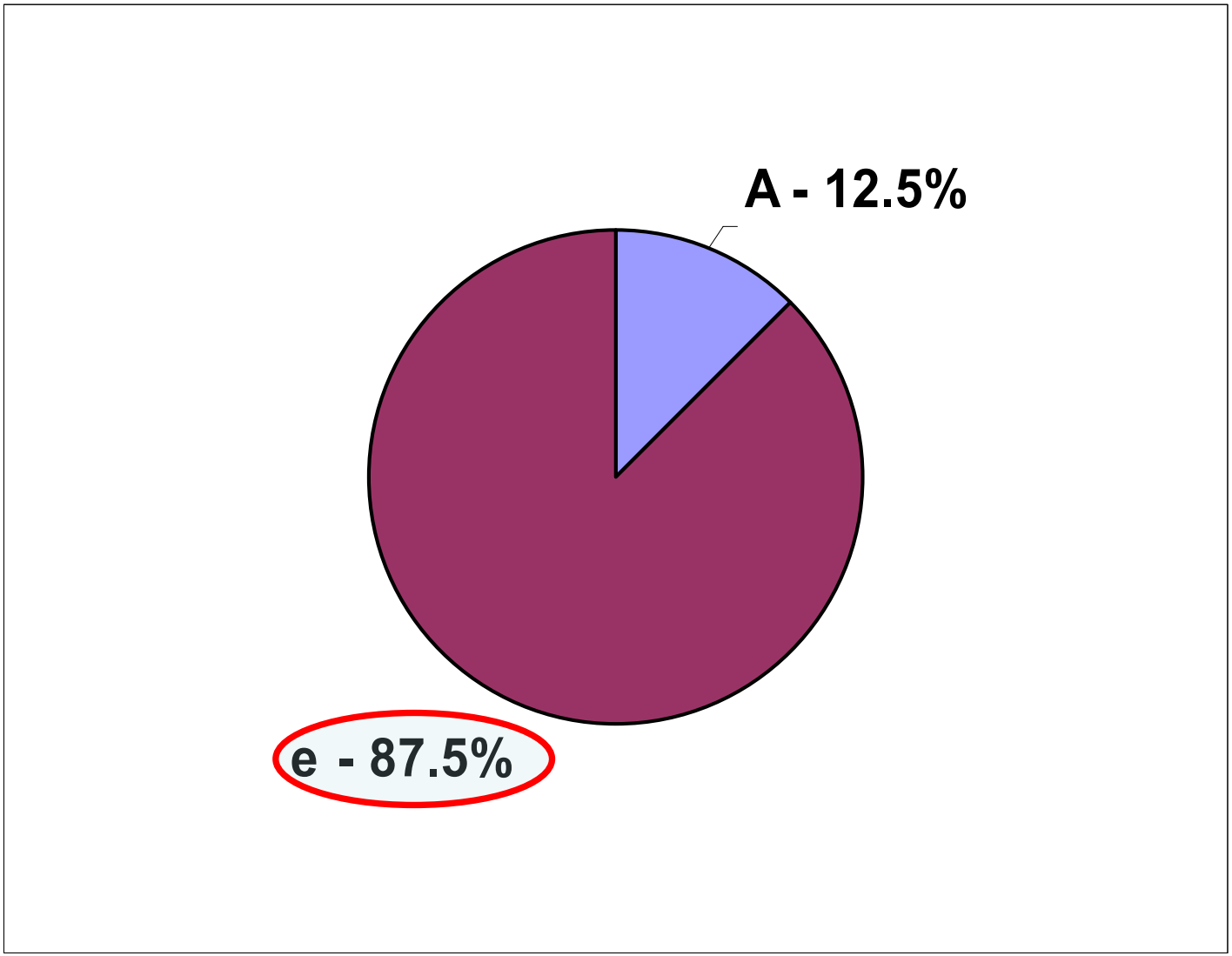
$$\eta^2\%_A = (SS_A/SS_{total})\cdot 100 = (12/96)\cdot 100 = 12.5\%$$

(для любой компоненты изменчивости)

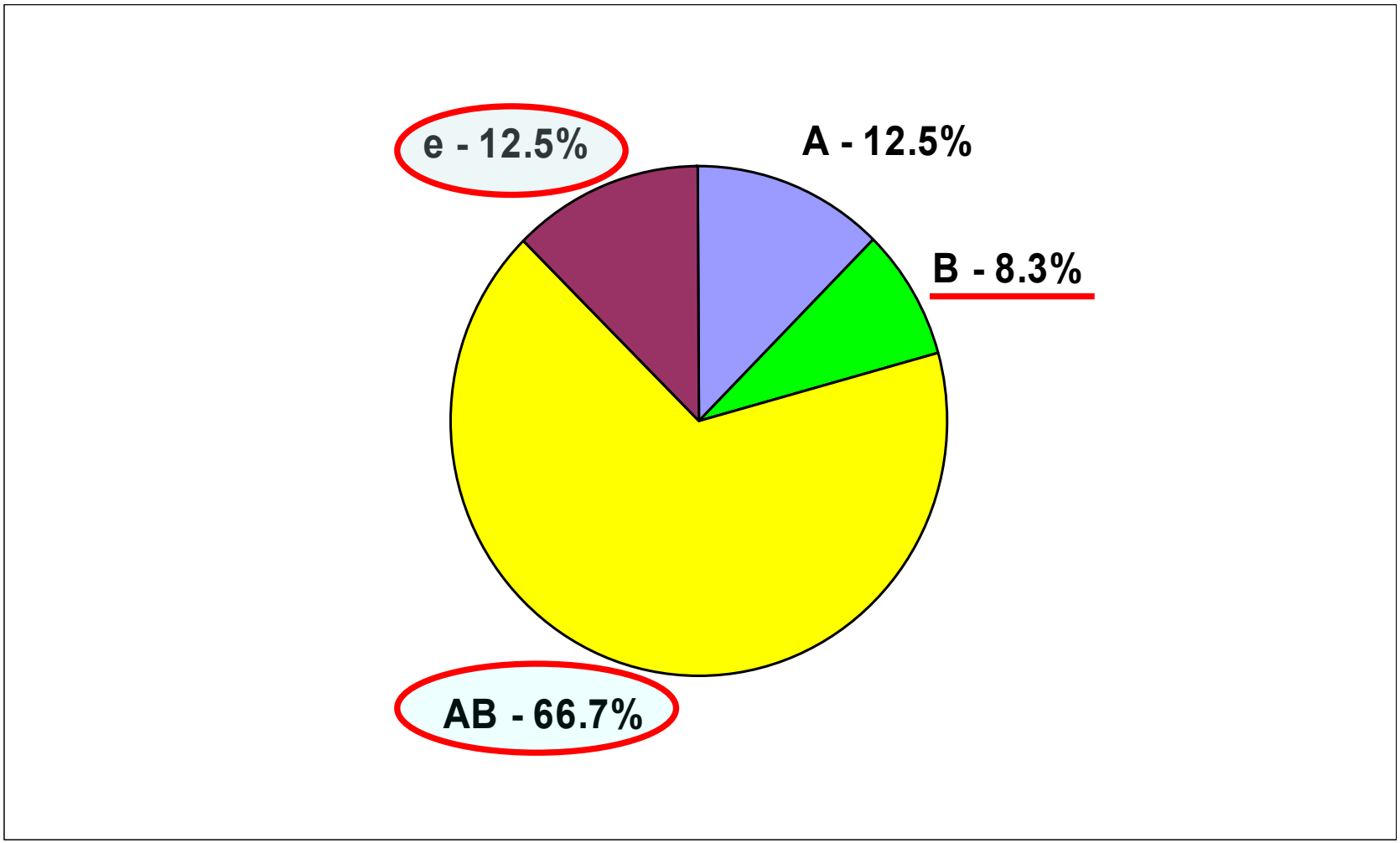
В русскоязычной литературе обычно используют названия

«доля» или «сила» влияния.

Сила влияния (однофакторная схема)



Сила влияния (двухфакторная схема)



Более точная оценка доли влияния (с учетом случайного варьирования групповых средних) может быть получена в таком виде:

$$\eta_x^{2} * = 1 - MS_e/MS_{tot} =$$

$$= [1 - (SS_e/SS_{tot})]\cdot((N-1)/(N-a))$$

Для нашего примера двухфакторного анализа

$$\eta^{2*}_x = 1 - (12/96)\cdot(17/12) = 0.8229$$

$$\eta^{2}_x = 84/96 = 0.8750$$

- Для *проверки* ***значимости*** корреляционного отношения в качестве грубого приближения
- $s(\eta^2_x) = (1 - \eta^2_x)/(n)^{1/2}$
- и затем

$$t = \eta^2_x / s(\eta^2_x)$$

- Однако более точная оценка может быть получена с использованием F-критерия:
- $F = \{\eta^2_x /(1 - \eta^2_x)\}\cdot\{df_2/df_1)$

- Что «добавляет» использование показателя «доля влияния»?

КОРЗУН

Владимир Михайлович

ПЛОТНОСТНО-ЗАВИСИМАЯ ТРАНСФОРМАЦИЯ СТРУКТУРЫ ПОПУЛЯЦИЙ И СООБЩЕСТВ НАСЕКОМЫХ

(НА ПРИМЕРЕ ДРОЗОФИЛЫ И БЛОХ)

03.00.16 – экология

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

доктора биологических наук

Иркутск – 2007

Дисперсионный анализ уровня плодовитости при различной плотности у линии популяции «Иноземцево-4» (трехфакторная схема) - *таблица автора*

Источник изменчивости	df	MS	F
Плотность	1	42193,10	689,99 ***
Линия	60	89,18	2,29 ***
Опыт	1	2380,32	61,11 ***
Взаимодействие линия - плотность	60	60,57	1,84 **
Взаимодействие линия - опыт	60	38,95	1,70***
Взаимодействие плотность - опыт	1	1392,98	42,34 ***
Взаимодействие линия - плотность - опыт	60	32,90	1,44*
Остаточная изменчивость	732	22,87	-

Можно «досчитать» по приведенным данным:

Источник изменчивости		η^2	P_0
Плотность	A	55.5	0.000000
Линия	B	7.0	0.000000
Опыт	C	3.1	0.000000
Взаим. линия – плотн.	AB	4.8	0.000181
Взаим. линия – опыт	BC	3.1	0.001092
Взаим. плотн. – опыт	AC	1.8	0.000000
Взаим. линия - плотн - опыт	ABC	2.6	0.019057
Остаточная изменчивость	e	22.0	

- Эффекты с примерно одинаковыми оценками вероятности P_0 могут иметь сильно различающиеся показатели доли влияния
- Оценка доли влияния и ее значимости не имеет прямой связи с «основными» результатами дисперсионного анализа (ДА)
- Даже при отрицательных результатах ДА можно использовать информацию о доле влияния исследованных факторов (по крайней мере для планирования дальнейших исследований)

- В англоязычных публикациях использование показателя «доли влияния» встречается реже
- Предлагается ограничить его применение определенными моделями (только для «случайных» факторов)
- Предлагаются разные способы оценки (и – вычисления)
- ОДНАКО «intraclass correlations» Фишера в отличие от «variance components» более традиционны, сопоставимы и проще интерпретируются.

Вернемся к нашему примеру…

- **Некоторые другие изменения в данных:**

  1) Если не проявляется влияние взаимодействия (специфика восприятия разных предметов девочками и мальчиками)

- После “перестановки” значений (соответствующих разным градациям фактора В в третьей градации по А, т.е. числа ошибок по литературе у мальчиков и девочек)

Данные двухфакторного анализа (после перестановки)

	B1	B2	Суммы (A)	Сумма квадратов
A1	1 2 3	5 6 7	24	
	Сумма : 6	сумма : 18		
A2	2 4 3	8 6 7	30	
	Сумма : 9	сумма : 21		
A3	3 4 5	7 8 9	36	
	Сумма : 12	сумма : 24		
Суммы (B)	27	63	90	546

Результаты двухфакторного дисперсионного анализа (после перестановки вариантов)

	SS	df	MS	F	$F_{0.05}$	P_0	$\eta^2\%$
X	84						87.5
A	12	2	6.0	6.0	3.88	.015625	12.5
B	72	1	72	72.0	4.75	.000002	75.0
AB	0	2	0				0
e	12	12	1				12.5
Total	96						

- Поскольку полная межгрупповая изменчивость совпадает в обоих случаях, мы получили лишь ее перераспределение между компонентами, соответствующими эффекту фактора В и взаимодействия АВ. При этом уже не проявляется "специфика" мальчиков и девочек в восприятии разных предметов (взаимодействие).

  За счет "передачи преимущества" в числе ошибок по литературе мальчики стали более достоверно отличаться от девочек

- **Другое изменение: увеличение объема выборок (при сохранении структуры данных)**

Увеличение объема групп (в 5 раз)

	SS	df	MS	F	P_0	η^2%	η^2*%
x	420						
A	60	2	30	42.0	0.000000	12.5	12.4
B	40	1	40	56.0	0.000000	8.3	8.2
AB	320	2	160	224.0	0.000000	66.7	66.1
Случ.	60	84	0.71429			12.5	13.3
Общая	480						

- **Изменения в результатах:**

  1) Резко уменьшилась вероятность P_0

  2) Поскольку число групп осталось прежним, а их объем увеличился, различия в оценках доли влияния «по Фишеру» и уточненной – уменьшились.

- «Общие» оценки в дисперсионном анализе не исключают возможности сравнения отдельных групп.
- Возможны два варианта такого сравнения:

  а) запланированное сравнение (оценка контрастов – для фиксированных факторов)

  б) post hoc сравнение
- Используется «обобщенная» оценка случайной изменчивости

“Post-hoc” сравнение отдельных групп (вариантов)

(подготовительные операции – по нашему примеру)

Разности между группами (№№ + средние)							
A	B	{№}	1 (2)	2 (6)	3 (3)	4 (7)	5 (8)	6 (4)
1	1	{1}						
1	2	{2}	4					
2	1	{3}	(1)	3				
2	2	{4}	5	(1)	4			
3	1	{5}	6	2	5	(1)		
3	2	{6}	2	2	(1)	3	4	

НСР (наименьшая существенная разница = Least Significant Difference) Фишера (P_0)						
A B		1	2	3	4	5
1 2 {2}		.00037				
2 1 {3}		.24417	.00318			
2 2 {4}		.00005	.24417	.00037		
3 1 {5}		.00001	.03062	.00005	.24417	
3 2 {6}		.03062	.03062	.24417	.00318	.00037

.24417 – P_0>0.05, .03062 – P_0<0.05, .00318 – P_0<0.01, .00037 – P_0<<0.001.

Тест Тьюки (тем больше отличается от предыдущего, чем больше сравниваемых групп)						
1 1 {1}						
1 2 {2}	.0039					
2 1 {3}	.8173	*.0295*				
2 2 {4}	.0007	.8173	.0039			
3 1 {5}	.0002	.2140	.0007	.8173		
3 2 {6}	.2140	.2140	.8173	*.0295*	.0039	

- тест НСР Фишера принимает как не случайную на первом уровне значимости (P_0<0.05) уже разность, равную 2,
- тест Тьюки = 3,
- тест Шеффе = 4.

ДРУГАЯ ФОРМА ПРЕДСТАВЛЕНИЯ =>Гомогенные группы»

N	Выборки		Средние	Группы			
	A	B		1	2	3	4
А) по тесту Фишера							
1	1	1	2	****			
3	2	1	3	****	****		
6	3	2	4		****		
2	1	2	6			****	
4	2	2	7			****	****
5	3	1	8				****
Б) по тесту Шеффе							
1	1	1	2	****			
3	2	1	3	****	****		
6	3	2	4	****	****	****	
2	1	2	6		****	****	****
4	2	2	7			****	****
5	3	1	8				****

Условия применимости (assumptions)

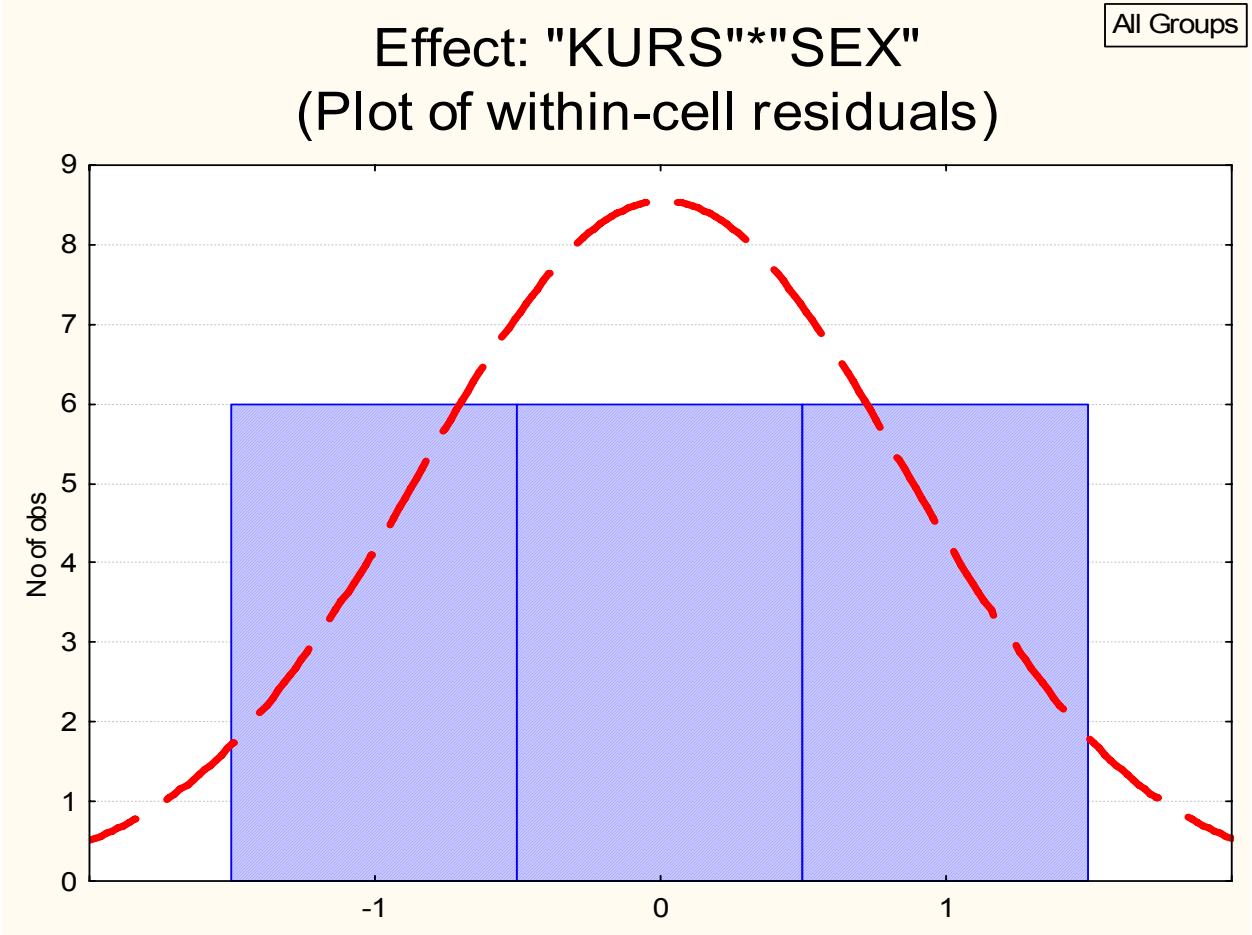
Зависимая переменная – интервальная
Значения факторов – дискретные (или порядковые; могут быть интервальными)

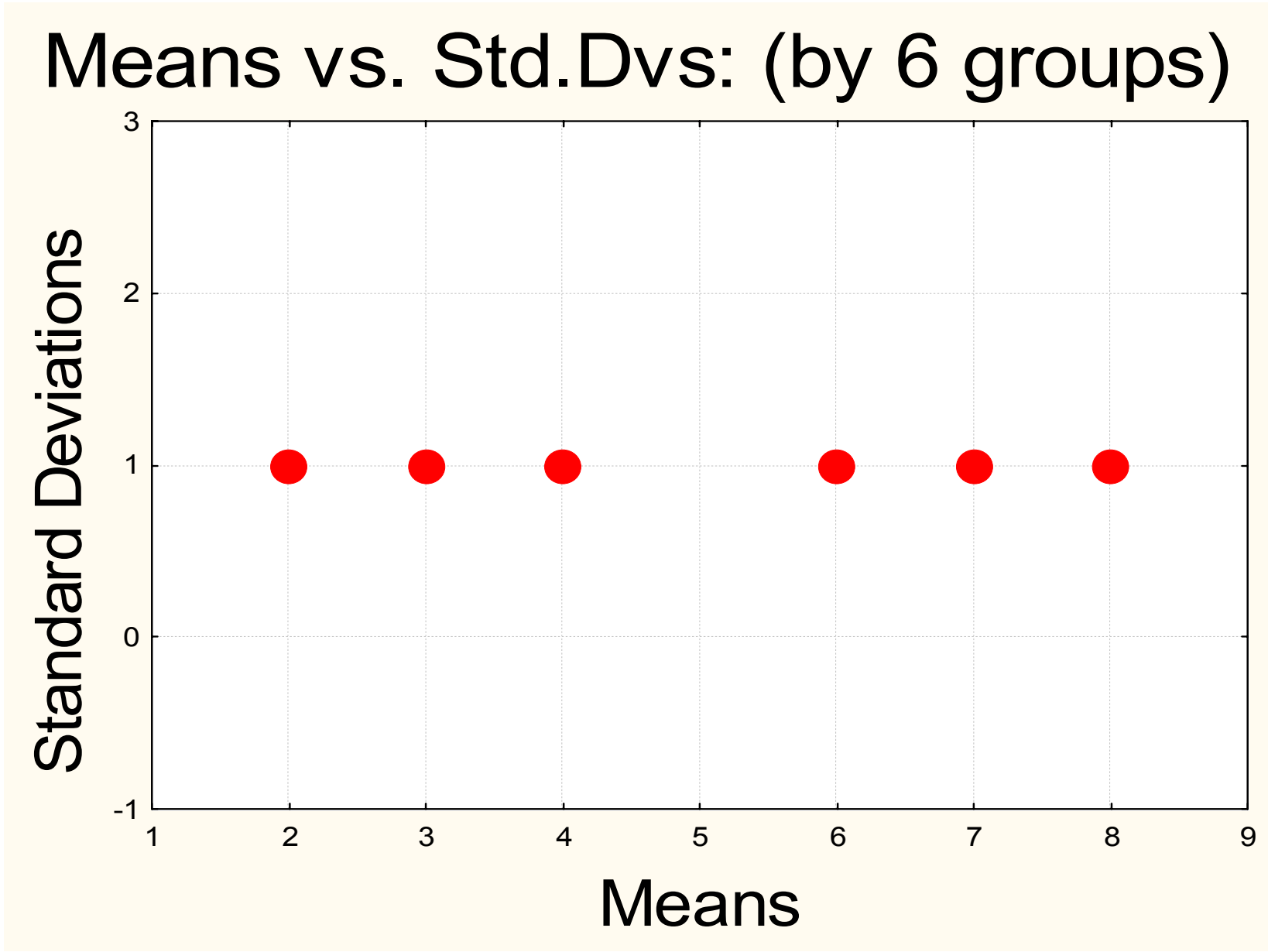
- **Нормальное распределение вариант в группах**
- **Равенство дисперсий в группах**
- Размеры групп приблизительно одинаковы
- Факторы независимы друг от друга (для многофакторной схемы)

«Нормальность» распределения

- Варьирование ***внутри групп*** должно соответствовать нормальному распределению.
- Т.е. «остаточное варьирование» должно иметь независимый и случайный характер

В нашем искусственном примере




Means vs. Std.Dvs: (by 6 groups)
Standard Deviations
3
2
1
0
-1
1
2
3
4
5
6
7
8
9
Means

На что именно и как влияет нарушение этого требования?

Г.Шеффе, 1980, с.396 (5 групп по 5 наблюдений)

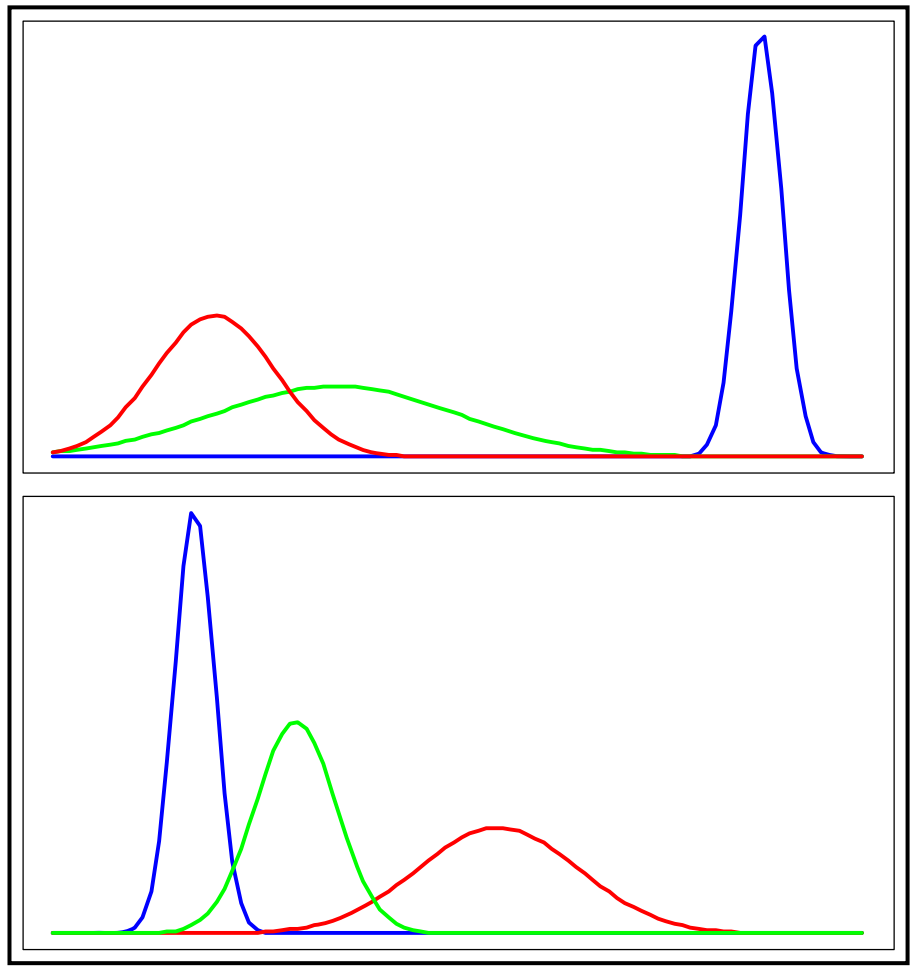
A - асимметрия, E- эксцесс

A	E				
	-1	-0.5	0	0.5	1
0	0.053	0.051	0.050	0.048	*
0.5	0.052	0.051	0.050	0.049	*
1.0	0.052	0.050	0.049	0.048	0.048

Гомогенность (гомоскедастичность) дисперсий

Гомогенность дисперсий
= гомоскедастичность =
= равенство дисперсий
в группах

Наихудший случай –
корреляция средних и
дисперсий



Нарушения «гомоскедастичности» (*Шеффе, 1980:401)*

Число групп	Отношение дисперсий в группах	Объемы групп	Общая численность	Вероятность ошибки первого рода
3	1:2:3	5 5 5	15	0.056
		3 9 3	15	0.056
		7 5 3	15	0.092
		3 5 7	15	0.040
	1:1:3	5 5 5	15	0.059
		7 5 3	15	0.110
		9 5 1	15	0.170
		1 5 9	15	0.013
5	1:1:1:1:3	5 5 5 5 5	25	0.074
		9 5 5 5 1	25	0.140
		1 5 5 5 9	25	0.025
7	1:1:1:1:1:1:7	3 3 3 3 3 3 3	21	0.120

Что же делать???

- Показанные в рассмотренных Шеффе примерах «нарушения» и, соответственно – возможность принятия некорректных решений (по крайней мере – для гипотез о различиях по средним) уменьшаются при увеличении объема выборок
- Наиболее опасны «нарушения» при близости к главной “условной границе”, соответствующей первому уровню значимости (P_0=0.05).

- Нарушение каждого из трех первых требований к данным для дисперсионного анализа может привести к ошибкам в оценках, поскольку в них мы используем значения критериев, построенных, исходя из предположения о соблюдении этих требований.
- Возможность ошибки усугубляется, если

1) одновременно нарушается не одно, а хотя бы два из требований

2) объем групп невелик (<30)

- Наиболее «доступные» операции:

1) выравнивание объемов групп.

2) трансформация данных

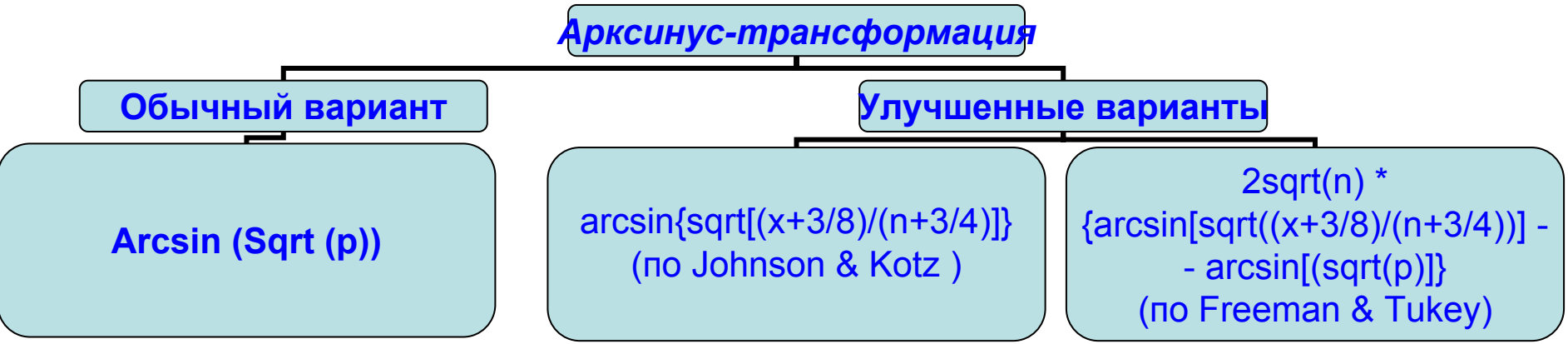
- **Рекомендуется**
- Логарифмирование в случаях, когда есть
  - Непрерывно распределенные переменные
  - Есть корреляция дисперсий и средних
  - У частотного распределения асимметрия вправо

Если	Вариант трансфор-мации
Нет нулевых значений	**Log (x)**
Есть нулевые значения	**Log (x + 1)**
Много значений 0<x<1; n ≥ 4	**Log (x * 10n)**

- Извлечение корня квадратного
  - Для дискретных переменных (распределение Пуассона)

Если	Вариант трансформации
Нет нулевых значений	Sqrt (x)
Есть нулевые значения	Sqrt (x + 1/2)

- Арксинус-трансформация (угловая трансформация)
  - *для процентов и долей*



Проверка соответствия основным требованиям

- Соответствие **нормальному распределению**

  1) по величине коэффициентов асимметрии и эксцесса

  2) по общему виду распределения

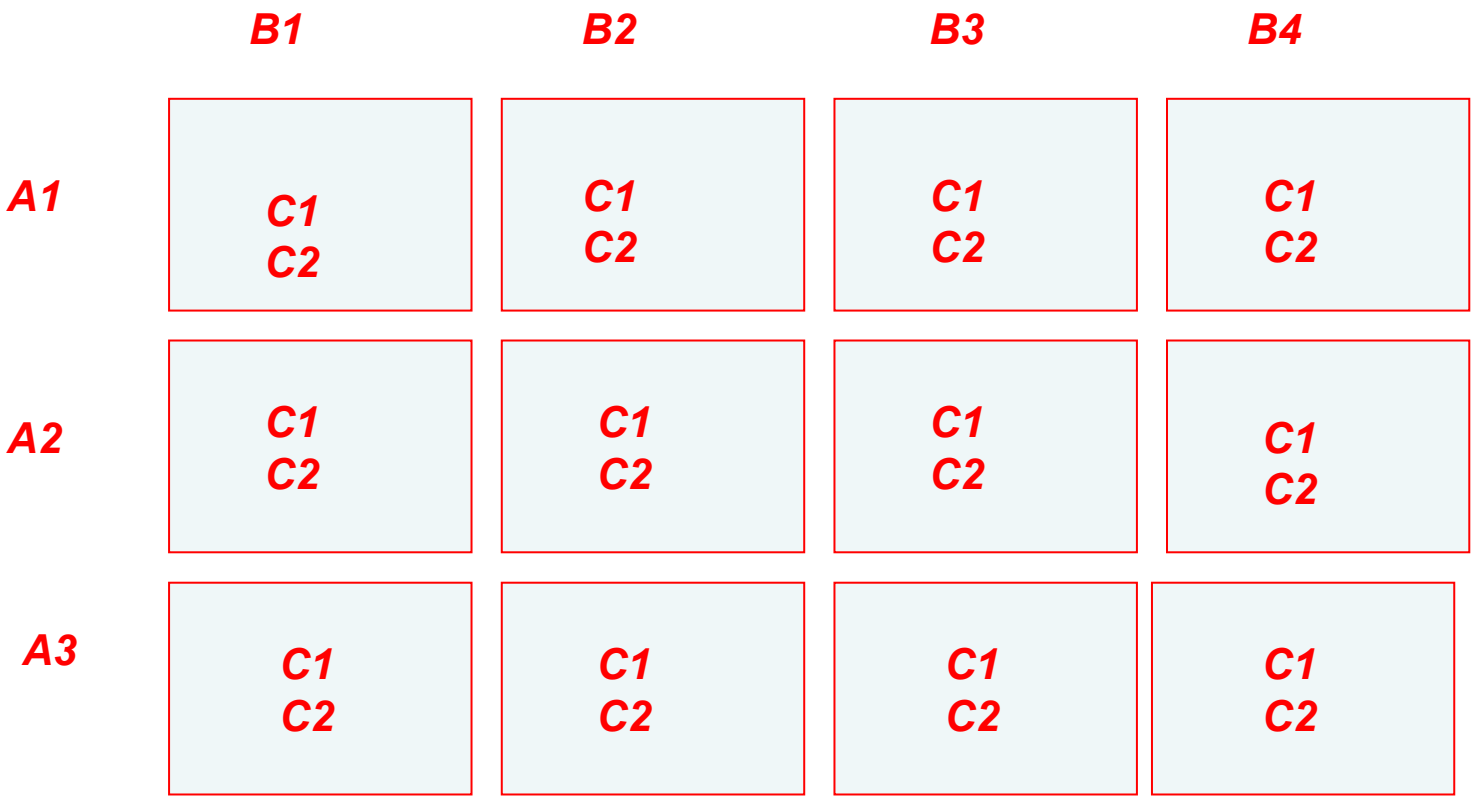
- Для анализа следует использовать только величины *остаточной (=случайной) изменчивости*

- **Равенство внутригрупповых дисперсий**
- **Методы оценки зачастую чувствительны к отклонениям от нормального распределения (например, критерий Бартлетта), а также к размерам выборок.**
- **Рекомендуется использовать графические методы анализа: например, сопоставление групповых дисперсий и средних**.

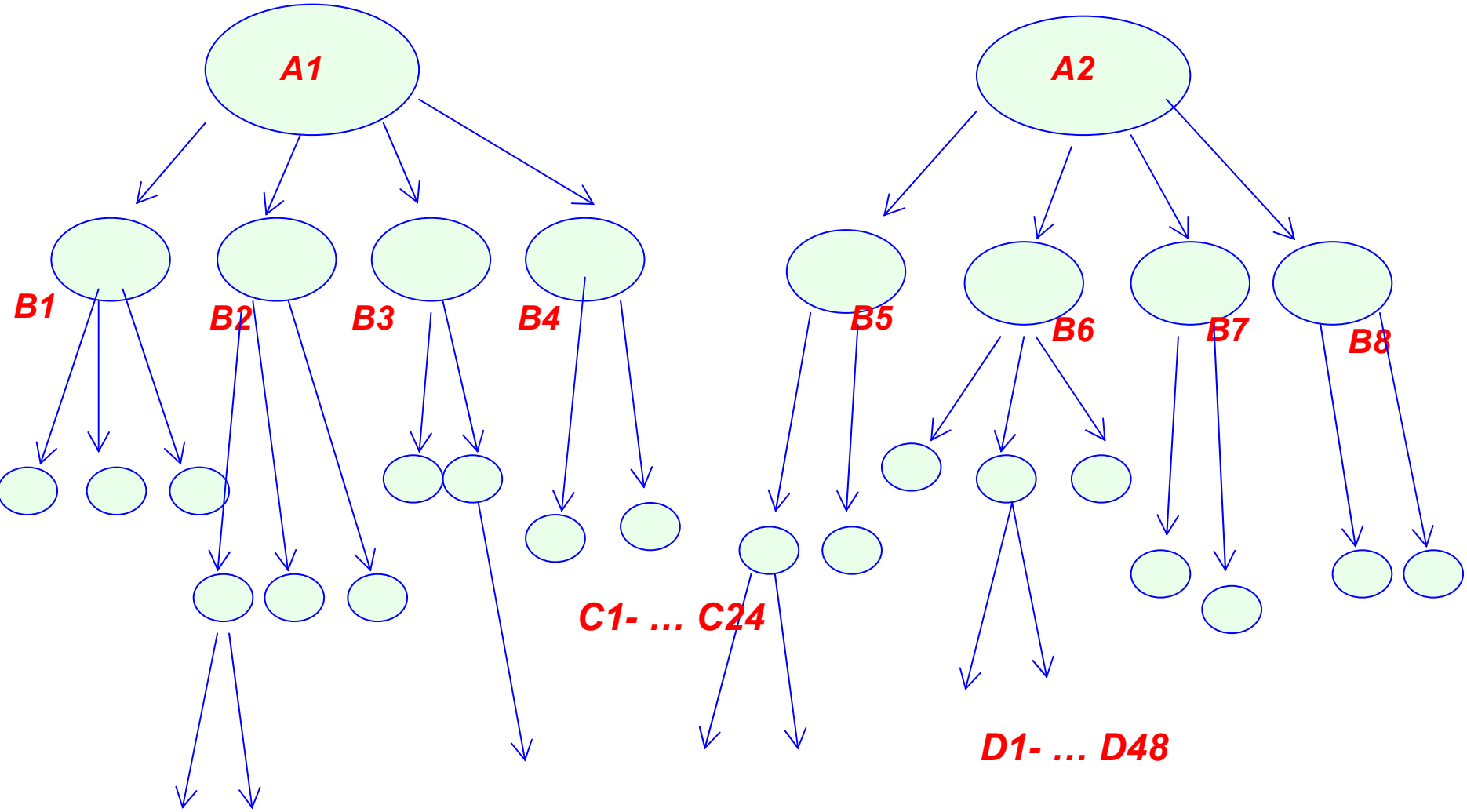
Разные схемы дисперсионного анализа

Кроме «обычной» схемы, в которой все исследуемые факторы образуют ортогональную систему, возможны также другие схемы многофакторного анализа:

- Иерархическая (= nested)
- С повторными измерениями (= repeated measures)



«Простая» = ортогональная схема дисперсионного анализа: градации каждого фактора (внутри другого) <u>сопоставимы</u> и могут быть объединены.



Иерархическая схема: соподчиненность факторов и несопоставимость их градаций.

Иерархическая схема

- Группировка по определенным факторам носит соподчиненный характер и, соответственно, градации одного фактора внутри другого не идентичны и поэтому не могут быть объединены.
- Примеры: а) виды – популяции – особи …

  б) родители – потомки (разных поколений)

Особенности анализа

- Вычисляемая по всем градациям фактора сумма квадратов отклонений включает оценку "чистого" влияния этого фактора и влияние всех вышележащих факторов.
- Невозможна оценка взаимодействия иерархических факторов.
- Определение значимости влияния каждого фактора - по отношению к среднему квадрату "нижележащего" (включенного в него) фактора.

Вычисления

- $SSt = Y - T$
- $SSx = A - T$
- $SS_A = A - B - T$
- $SS_B = B - C - T\ldots$

- $SSe = Y - A$

- $F_A = MS_A/MS_B$
- $F_B = MS_B/MS_C$
- $F_C = MS_C/MS_e$

ПРИМЕР: Снедекор. 1961. с.252; табл. 1.5.1

Растения (A)	Листья (B(A))	Пробы (e)
1	1	3.28 3.09
	2	3.52 3.48
	3	2.88 2.80
2	1	2.46 2.44
	2	1.87 1.92
	3	2.19 2.19
3	1	2.77 2.66
	2	3.74 3.44
	3	2.55 2.55
4	1	3.78 3.87
	2	4.07 4.12
	3	3.31 3.31

Подготовительные вычисления (суммы по градациям факторов):

Общая сумма: **72.29**, сумма квадратов **228.0139.**

"Поправка"

C = $(72.29)^2/24$ = 217.7435

Растения	Листья
19.05	6.37
	7.00
	5.68
13.07	4.90
	3.79
	4.38
17.71	5.43
	7.18
	5.10
22.46	7.65
	8.19
	6.62

Вычисления:

- $SSt = 228.0139 – 217.7435 = \underline{10.2704}$
- $SS_A = 1/6 \cdot(19.05^2 + 13.07^2 + 17.71^2 + 22.46^2) – 217.7435 =$

$= 225.30385 – 217.7435 = \underline{7.5604}$

- $SS_{B(A)} = 1/2 \cdot(6.37^2 + 7.00^2 + \ldots \ldots + 6.62^2) - 225.30385 =$

$= 227.93405 – 225.30385 = \underline{2.6302}$

- $SSe = 228.0139 – 227.93405 = \underline{0.07985}$

- Число степеней свободы
- для A = a-1 = 4 – 1 = 3
- для B(A) = a(b-1) = 4(3-1) = 8
- для e = ab(n-1) = 12(2-1) = 12

Результаты иерархического дисперсионного анализа

Изменчивость	SS	df	MS	F	P_0	$\eta^2\%$	$\eta^2(-e)\%$
Растения (A)	7.56035	3	2.520115	7.665	0.009725	73.6	73.1
Листья (B(A))	2.63020	8	0.328775	49.41	0.000000	25.6	25.4
Пробы	0.07985	12	0.006654			0.8	1.5
Общая	10.2704					100	100

Иная форма заключительной таблицы для иерархического дисперсионного анализа:

	Эффект		Случайная		F	P_0	$\eta^2\%$
	df	MS	df	MS			
Растения	3	2.520115	8	0.328775	7.665167	0.009725	73.6
Листья	8	0.328775	12	0.006654	49.408890	0.000000	25.6
Пробы (e)	12	0.006654					0.8

- Обратите внимание:
- Наибольшая доля изменчивости определяется особенностями растений (73% при $0.01>P_0>0.001$).
- Различия листьев на растении (при еще более высокой достоверности оценки $P_0<<0.001$) составляют лишь 25%
- Случайная изменчивость (повторные пробы с одного и того же листа) – только 1.5%

Какие еще можно сделать выводы – на дальнейшее:

- **Точность химического анализа высока и пробы с одного листа настолько однородны, что вполне можно было бы ограничиться не двумя, а одним образцом.**
- **Кроме того, для надежной оценки можно, не увеличивая общее число проб, брать не три, а два или даже по одному листу с каждого растения[1], но больше растений.**
- **В особенности это будет важно, если мы захотим сравнить разные группы растений (например, сорта) или растения, выращенные в разных условиях.**
- **В этом случае, если для анализа будет взято небольшое число растений, мы можем принять случайные различия между выборками (из-за попадания в них сильно различающихся растений) за достоверные.**

- [1] Разумеется, при этом желательно было бы определить, с чем именно связаны различия по содержанию кальция между листьями с одного растения

Если действительно нужно сравнить по содержанию кальция два разных сорта турнепса:

1	*По 4 растения (+1)*							
	Эффект		Случайная		F	p-level	SS	%%
	df	MS	df	MS				
Сорта	1	12.0000	6	2.5201	4.7617	0.071850	12.0000	36.9
Раст.	6	2.5201	16	0.3288	7.6652	0.000527	15.1207	46.5
Листья	16	0.3288	24	0.0067	49.4089	0.000000	5.2604	16.1
Пробы	24	0.0067					0.1597	0.5
Общая							32.5408	

Результаты анализа показывают, что при таком уровне различий между «сортами» (+1) они остаются недоказанными (P_0>0.05). Поскольку мы просто продублировали все данные, соотношение остальных источников изменчивости (растения – листья – пробы) осталось примерно таким же (по величине доли изменчивости).

1a	*Если не учитывать иерархические факторы*							
	Эффект		Случайная		F	P	SS	%%
	df	MS	df	MS				
Сорта	1	12.0000	46	0.4465	26.873	0.000005	12.0000	36.9
Случ.	46	0.4465					20.5408	63.1
Общая							32.5408	

А если различия сортов выражены сильнее?
Примем разницу между нашими предполагаемыми "сортами“ равной +2

	Эффект		Случайная		F	p-level	SS	%%
	df	MS	df	MS				
2	*По 4 растения (+2)*							
Сорт	1	48.0000	6	2.5201	19.0467	0.004749	48.0000	70.0
Раст.	6	2.5201	16	0.3288	7.6652	0.000527	15.1207	22.1
Листья	16	0.3288	24	0.0067	49.4089	0.000000	5.2604	7.7
Пробы	24	0.0067					0.1597	0.2
Общая							68.5408	

А если просто увеличить число растений в пробе?

	Эффект		Случайная		F	p-level	SS	%%
	df	MS	df	MS				
3	По 8 растений (+1)							
Сорта	1	24.0000	14	2.1601	11.1106	0.004925	24.0000	36.9
Раст.	14	2.1601	32	0.3288	6.5701	0.000005	30.2414	46.5
Листья	32	0.3288	48	0.0067	49.4089	0.000000	10.5208	16.2
Пробы	48	0.0067					0.3194	0.5
Общая							65.0816	

А что получится, если брать всего по 2 растения?

	Эффект		Случайная		F	p-level	SS	%%
	df	MS	df	MS				
4а	*По 2 растения (выбор «разных» растений) (+1)*							
Сорта	1	6.0000	2	7.3477	0.8166	0.461558	6.0000	26.6
Раст.	2	7.3477	8	0.2362	31.1122	0.000168	14.6953	65.0
Листья	8	0.2362	12	0.0011	209.9259	0.000000	1.8893	8.4
Пробы	12	0.0011					0.0135	0.1
Общая							22.5982	
4б	*По 2 растения (выбор «типичных» растений) (+1)*							
Сорта	1	6.0000	2	0.1496	40.0980	0.024043	6.0000	61.1
Раст.	2	0.1496	8	0.4214	0.3551	0.711619	0.2993	3.0
Листья	8	0.4214	12	0.0122	34.5869	0.000000	3.3711	34.3
Пробы	12	0.0122					0.1462	1.5
Общая							9.8165	

Иногда исследователь заявляет, что в выборку взяты «только типичные» особи; иногда в качестве условия обнаружения «эффекта воздействия» требуют, например, собирать только «средние листья» с дерева…

В таком случае умышленно (или из-за недопонимания) занижается внутривыборочная изменчивость.

«Преобразование» иерархического анализа в «обычный»:

выбор конкретных градаций (контрастов)

Схема с повторными измерениями

- Аналог «связанных» (зависимых) выборок = оцениваются разности между повторными измерениями
- Можно использовать «величину реакции», т.е. разностей, а не абсолютных величин
- Оценка «индивидуальной» изменчивости как самостоятельной компоненты изменчивости

Основные модели дисперсионного анализа (фиксированные и случайные эффекты)

	Модель I фиксированные эффекты	***Модель II*** случайные эффекты
Градации фактора	Строго определены	Выбраны случайно из множества возможных
Повторение исследования	Возможно (точное)	Невозможно
Использование результатов	Только на изученный интервал (интерполяция)	На весь возможный диапазон (экстраполяция)

F-критерий в разных моделях

(Для однофакторного анализа различий НЕТ!!!)

	A и B		A – фикси-рованное, B случайное
	Фиксированные	Случайные	
A	$\mathbf{mS_A/mS_e}$	$\mathbf{mS_A/mS_{AB}}$	$\mathbf{mS_A/mS_e}$
B	$\mathbf{mS_B/mS_e}$	$\mathbf{mS_B/mS_{AB}}$	$\mathbf{mS_B/mS_{AB}}$
AB	$\mathbf{MS_{AB}/mS_e}$	$\mathbf{mS_{AB}/mS_e}$	$\mathbf{mS_{AB}/mS_e}$

- А если значимость взаимодействия – *не доказана?*
- Считается, что в этом случае невозможно получить оценку значимости самих факторов…
- В некоторых случаях предлагают для этого суммировать оценку дисперсии взаимодействия и внутригрупповую…

Планирование

Планирование

- соотношение числа вариантов и объема выборок (Гинзбург, 1973, с.157)

	N (общее число измерений)											
η^2	8	12	30	40	60	80	100	120	160	200	240	300
0.1	2	2	3	4	6	8	10	12	16	20	24	30
0.2	2	2	5	8	10	16	20	20	32	40	40	50
0.3	2	2	6	10	15	20	25	30	40	50	60	75
0.4	2	3	10	10	15	20	25	30	40	50	60	75
0.5	2	4	10	10	20	20	25	40	40	50	80	100
0.6	4	6	10	20	20	40	50	50	80	100	80	100
0.7	4	6	15	20	30	40	50	60	80	100	80	150
0.8	4	6	15	20	30	40	50	60	80	100	80	150
0.9	4	6	15	20	30	40	50	60	80	100	120	150

Следующая проблема: общая структура исследования.

Как подбирать градации (варианты) по каждому фактору? Каково должно быть соотношение градаций (и их числа) в многофакторном исследовании?

Такая задача весьма актуальна: ведь даже при трех факторах и числе градаций по каждому a=b=c=3 понадобится 3*3*3=27 разных вариантов! Ну, а с добавлением еще одного фактора их станет уже 81... Ведь в соответствии с требованием ортогональности эксперимента все градации каждого фактора должны быть представлены в каждой из градаций всех остальных.

- **«Дробные реплики»**

Полный трехфакторный план

Варианты	Факторы		
	A	B	C
1	+	+	+
2	+	+	-
3	+	-	+
4	+	-	-
5	-	+	+
6	-	+	-
7	-	-	+
8	-	-	-

То же – со взаимодействием

Варианты	Факторы			Взаимодействие
	A	B	C	ABC
1	+	+	+	+
2	+	+	-	-
3	+	-	+	-
4	+	-	-	+
5	-	+	+	-
6	-	+	-	+
7	-	-	+	+
8	-	-	-	-

Знаки для четвертого фактора определяем в соответствии со знаками взаимодействия

Варианты	Факторы			Взаимодействие	Фактор (новый)
	A	B	C	ABC	D
1	+	+	+	+	+
2	+	+	-	-	-
3	+	-	+	-	-
4	+	-	-	+	+
5	-	+	+	-	-
6	-	+	-	+	+
7	-	-	+	+	+
8	-	-	-	-	-

«Полу-реплика» для трех факторов
(четыре варианта вместо восьми)

	A+	**A-**
B+	**C+**	**C-**
B-	**C-**	**C+**

«Полу-реплика»

для четырех факторов (8 вариантов)

	A+		A –	
B+	C+ D+	C – D –	C – D+	C+ D –
B –	C – D+	C+ D –	C+ D+	C – D –

- Применение полу-реплики приводит к потере возможности оценки влияния взаимодействия «предыдущего» порядка:

  если это 4-факторный анализ, то мы не можем оценить по полу-реплике тройные и «четверные» взаимодействия

- **Аналогичные приемы используются для тех случаев, когда число градаций факторов больше двух. Сокращение общего числа вариантов в такой ситуации достигается за счет рандомизированного размещения вариантов в так называемых *латинских и греко-латинских квадратах*.**
- **В тех случаях, когда условия эксперимента или сбора материала не позволяют, например, осуществить полную программу в один и тот же период времени или на одном и том же участке, для сравнимости результатов применяют так называемые *"блоковые схемы".***

- ***Различия попарного сравнения и дисперсионного анализа* (ПС – ДАн)**

В ДАн все основные суждения получаем по *обобщенным* оценкам, в ПС – по оценкам конкретных выборок. При этом изменяются величины SS, df и, соответственно – $P_{0.}$

Кроме статистической значимости в ДАн можем получить оценку *«доли влияния»* каждого фактора и их взаимодействия

Вероятность случайности наблюдаемых различий (P_0) в ДАн определяется с учетом числа групп (более *корректная* оценка!)

Таким образом дисперсионный анализ по сравнению с попарными сравнениями является

А) более информативным

Б) более корректным

Советы

- Планируя исследование, тщательно обдумайте общую схему получения данных: главное – обеспечить сопоставимость влияния интересующих вас факторов.
- При получении данных для анализа следует стремиться к равным объемам групп (особенно ввиду возможных нарушений и сложностей).

- Если нет предварительных сведений о влиянии фактора, который предполагается изучать, лучше для начального этапа исследования выбрать наиболее контрастные варианты из всех возможных
- Гомогенность вариансс – очень важна! При нарушении этого требования ситуацию сильно осложняет неравенство групп и/или согласованность изменений по группам средних и стандартных отклонений.

- Негомогенность внутригрупповых дисперсий лучше оценивать по графикам: сопоставление средних и стандартных отклонений, остаточной изменчивости и - средних по группам.
- Обнаруженная в ваших данных негомогенность дисперсий осложняет анализ, но (в том случае, если это не результат ошибок в записях и т.п.) она же дает интересный материал для размышлений.

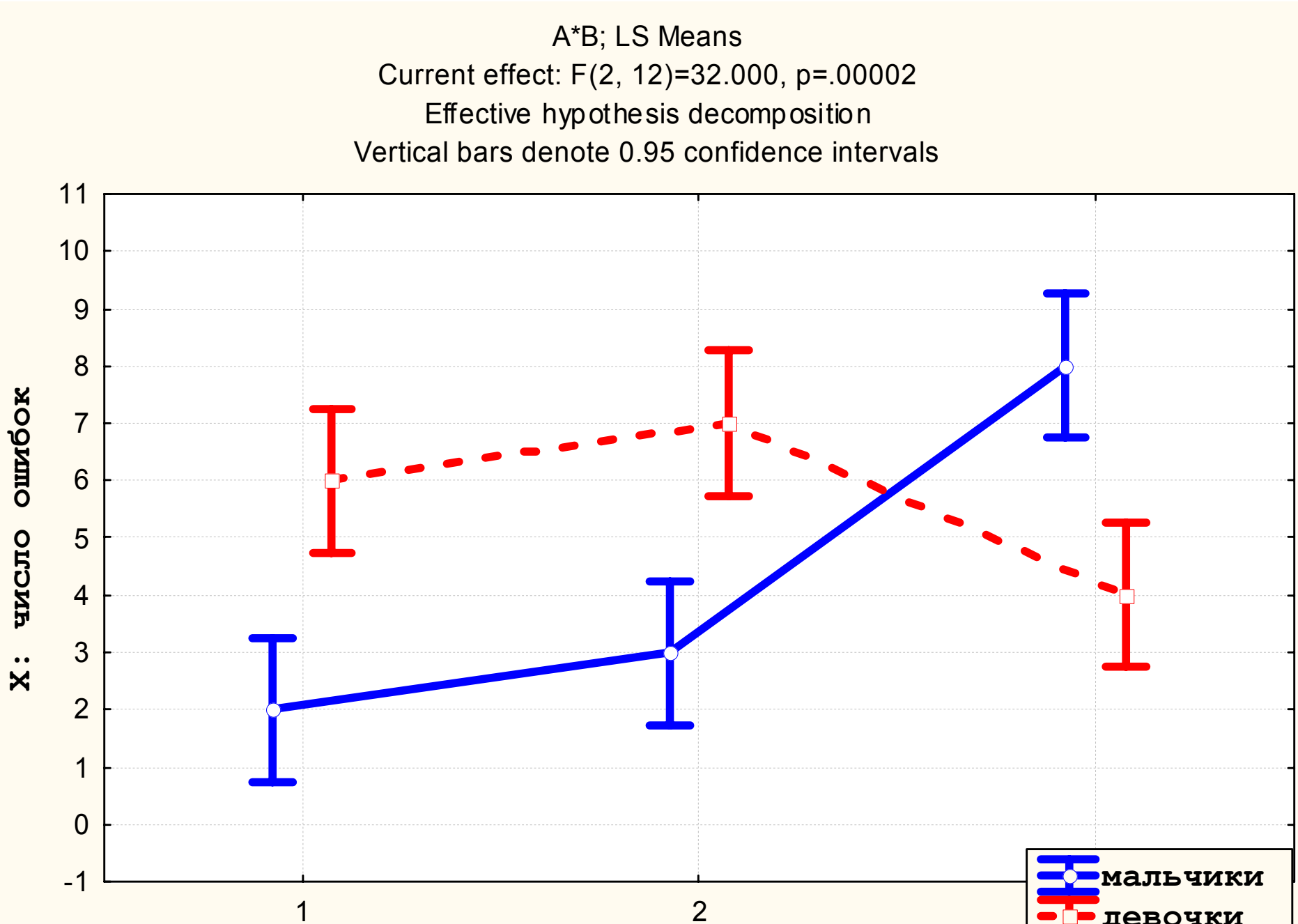
- Трансформация данных обычно применяется при асимметрии распределения; чаще всего – используют логарифмы и корень квадратный. Она полезна также для выравнивания групповых варианс.
- При умеренных отклонениях от ограничений к данным следует критично относиться к результатам, близким к маргинальным по значимости. (0.05, 0.01)
- При post hoc сравнениях рекомендуется применять тест Тьюки

- Наряду с таблицами обязательно следует использовать графическое представление результатов.
- «Формальное» доказательство значимости различий не должно противоречить известным биологическим закономерностям.

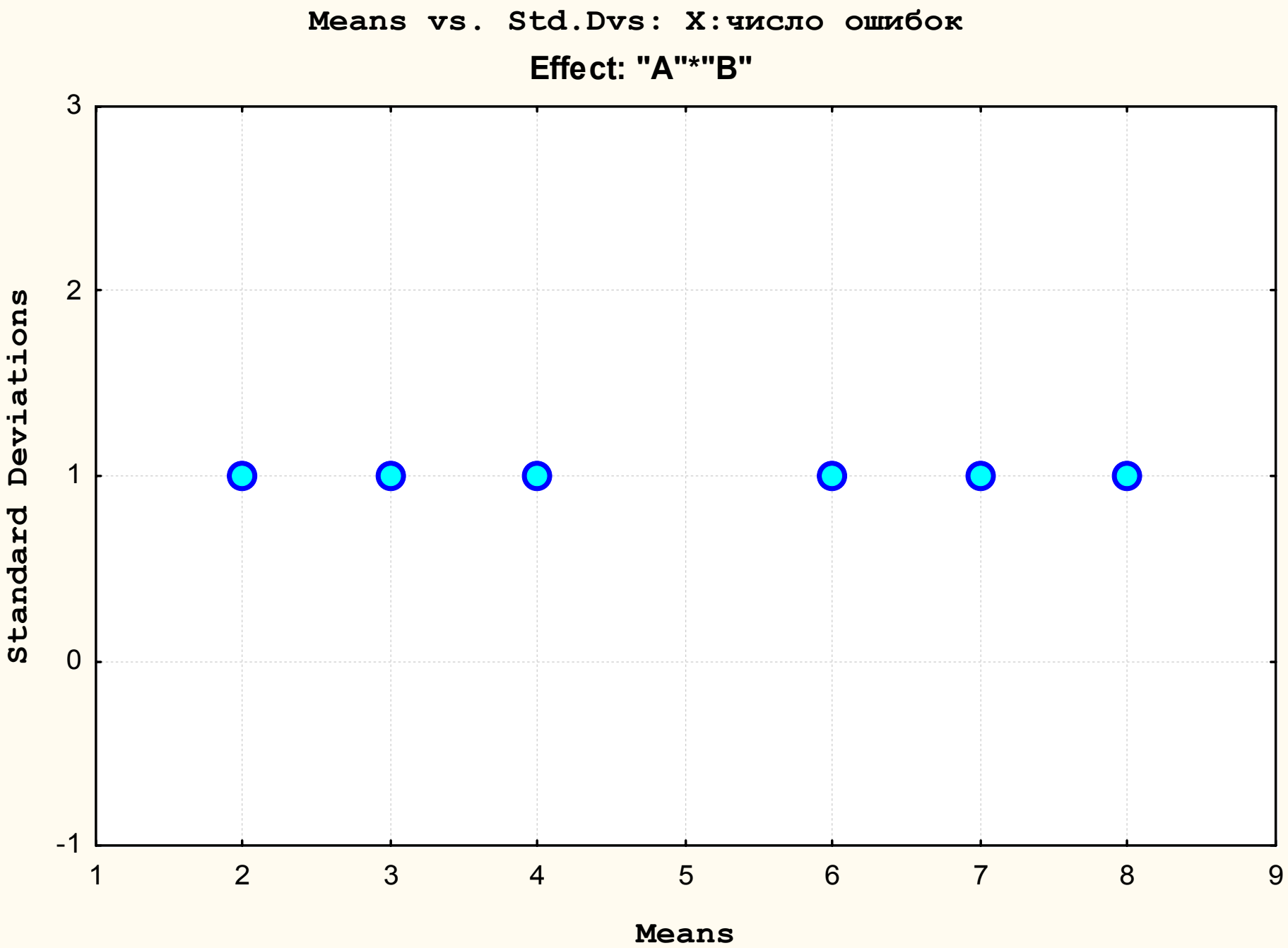
«Графические» результаты

ЕЩЕ РАЗ: КАРТИНКИ НАГЛЯДНЕЕ ТАБЛИЦ!!!

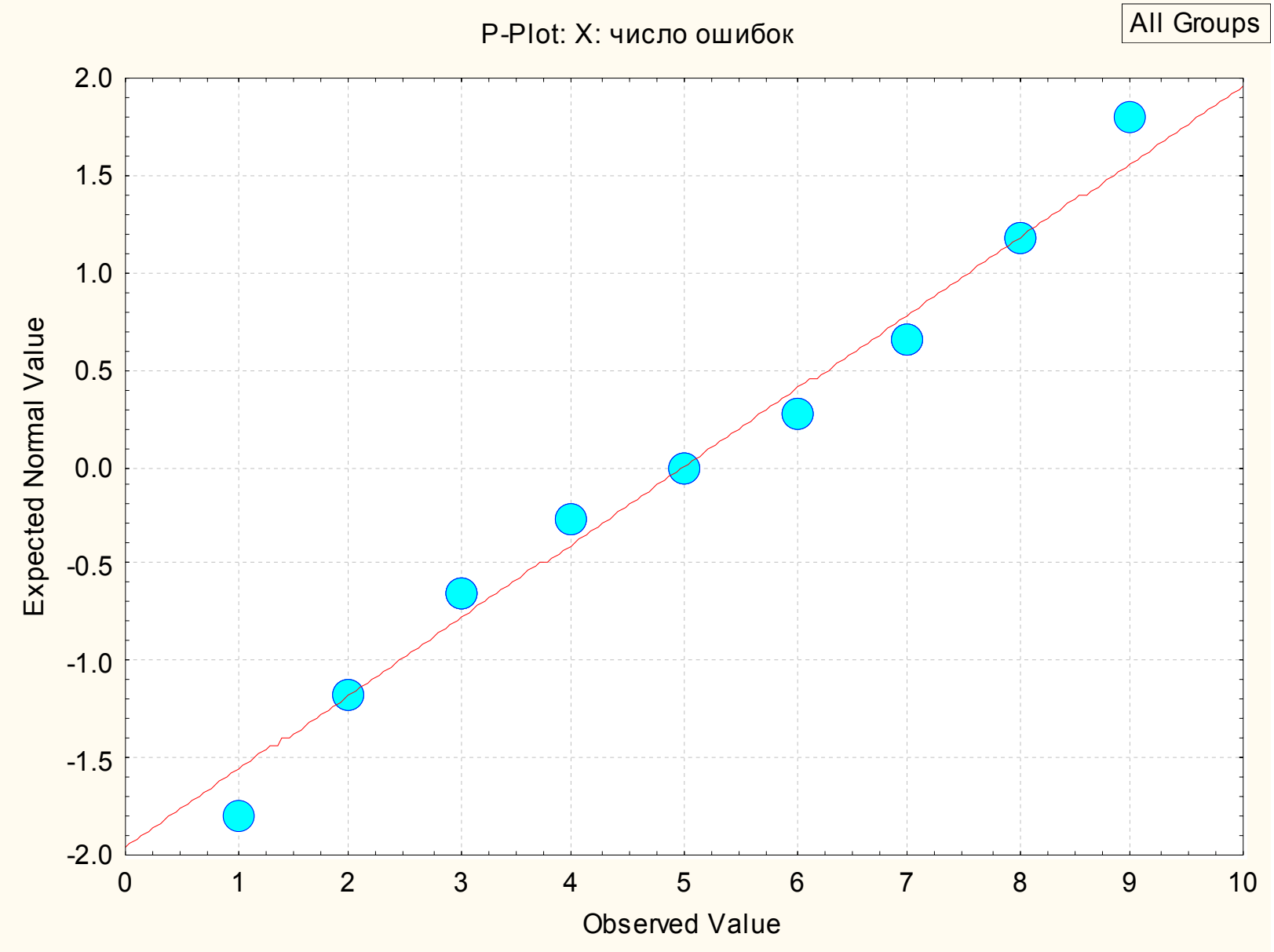
СОПОСТАВЛЕНИЕ СРЕДНИХ ПО ГРУППАМ



СОПОСТАВЛЕНИЕ СРЕДНИХ И ДИСПЕРВИЙ ПО ГРУППАМ



Характер распределения



Несколько слов про *MANOVA*...
= ***Многомерный Дисперсионный анализ***

Результат выражается в

1) значении многомерного критерия (Wilks – λ и др.) – показывает скоррелированность всех зависимых переменных с фактором (факторами)

2) вероятности – P_0

При этом можно одновременно получить и результаты для каждой зависимой переменной отдельно.

Примеры

М.Н. Олонова (Томский университет)

Близкие виды мятликов: Poa nemoralis, P. palustris

Экологический трансект (склон холма – берег реки – луг - опушка леса = 5 групп из 15 выборок + 2 «типовые» для видов; Всего 441 растение.)

21 измеренных признаков + 16 «индексов»;

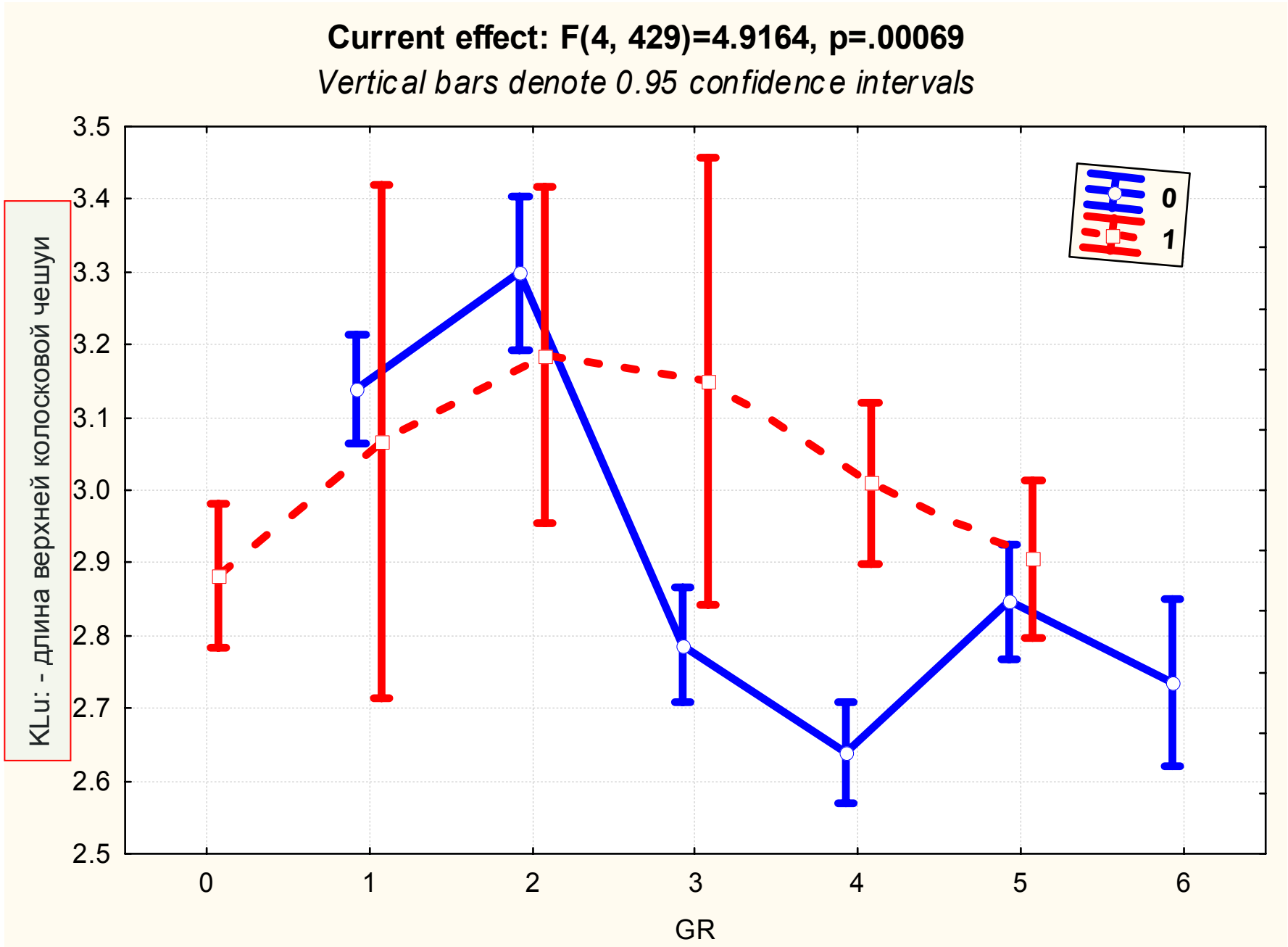
Multivariate Tests of Significance (факторы: Q1 «видовой» признак, GR – участки трансекта)

	Test Wilks	F	Effect df	Error df	P
"Q1"	1.000000		0		
GR	0.463144	33.32782	12	852	0.000000
"Q1"* GR	0.954891	1.65769	12	852	0.071392

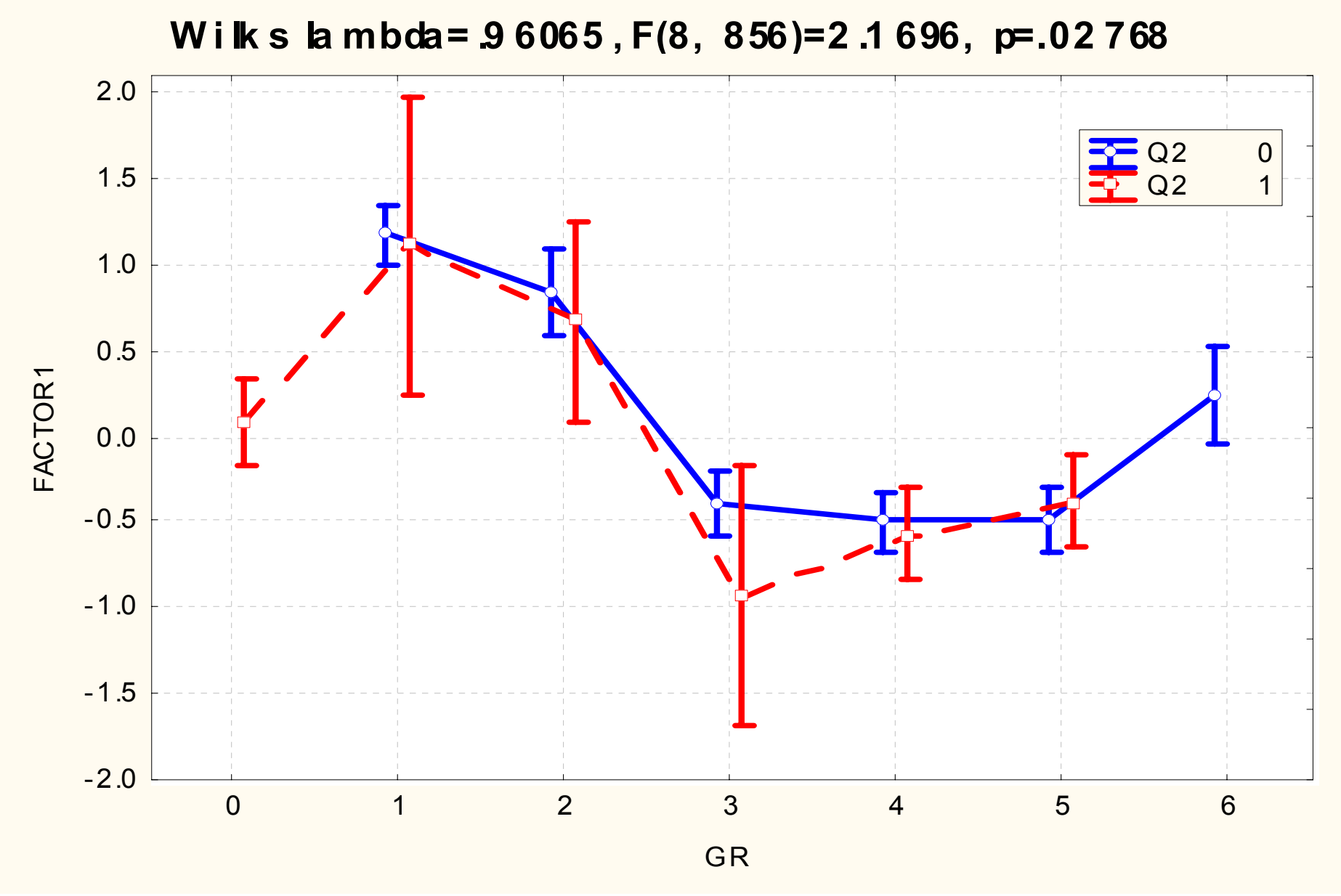
Univariate Tests of Significance
(один из признаков)

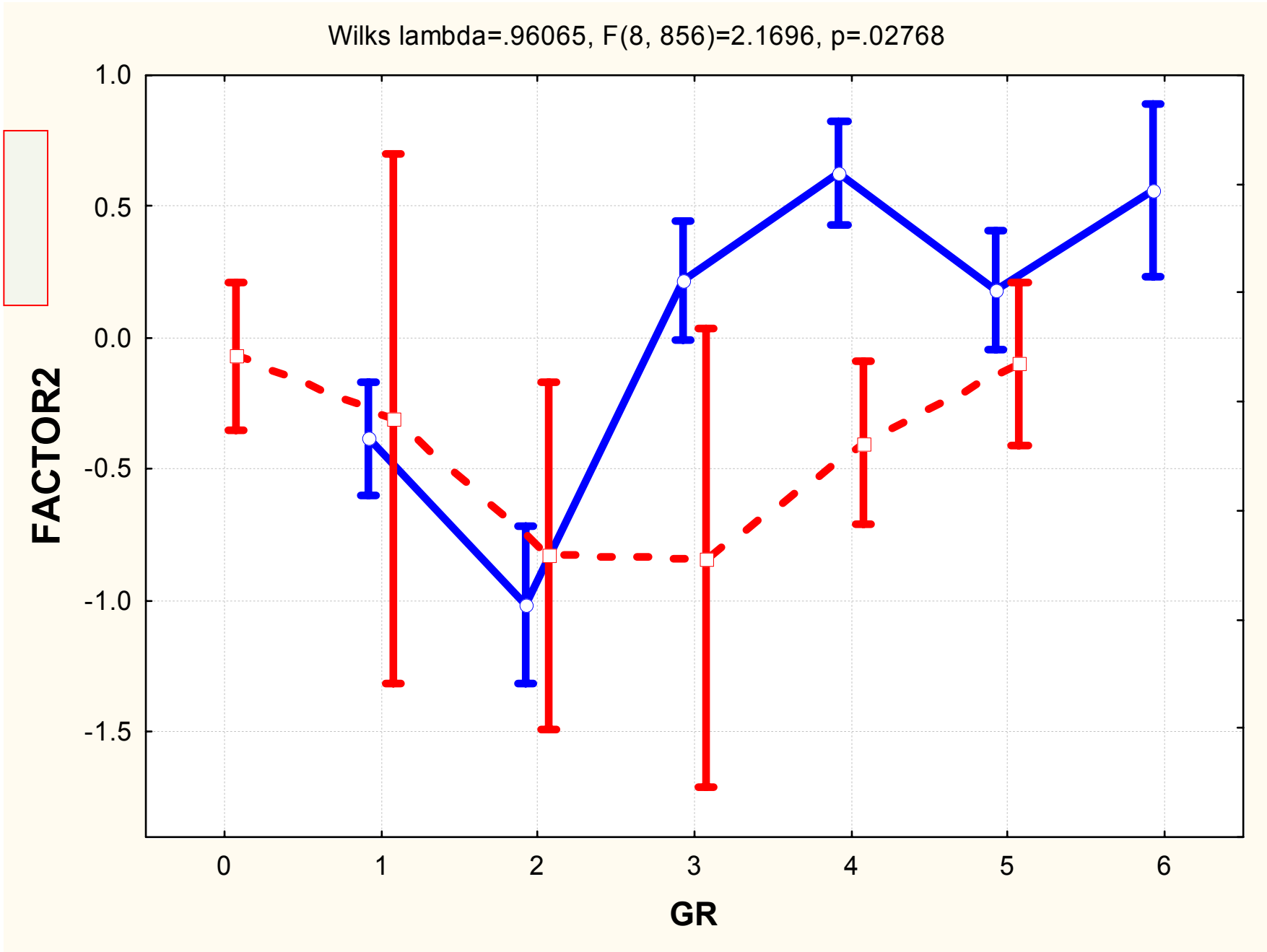
KLu	df	SS	MS	F	p
"Q1"	0				
GR	2	5.559	2.779	26.942	0.000000
"Q1"*GR	2	0.927	0.463	4.491	0.011741
Error	431	44.461	0.103		

H высота	df	SS	MS	F	Po	%%
GR	2	77533.5	38766.7	118.3	0.0000	26.0
"Q1"*GR	2	139.9	70.0	0.2	0.8079	0.0
Error	431	141247.3	327.7			47.4
L лист						
GR		953.8	476.9	83.1	0.0000	19.8
"Q1"*GR		31.4	15.7	2.7	0.0660	0.7
Error		2474.6	5.7			51.3
M метелка						
GR		710.1	355.0	35.9	0.0000	11.5
"Q1"*GR		35.0	17.5	1.8	0.1722	0.6
Error		4268.0	9.9			69.1
Total		4824.0				
KL кол.чешуя						
GR		5.6	2.8	26.9	0.0000	9.4
"Q1"*GR		0.9	0.5	4.5	0.0117	1.6
Error		44.5	0.1			75.2

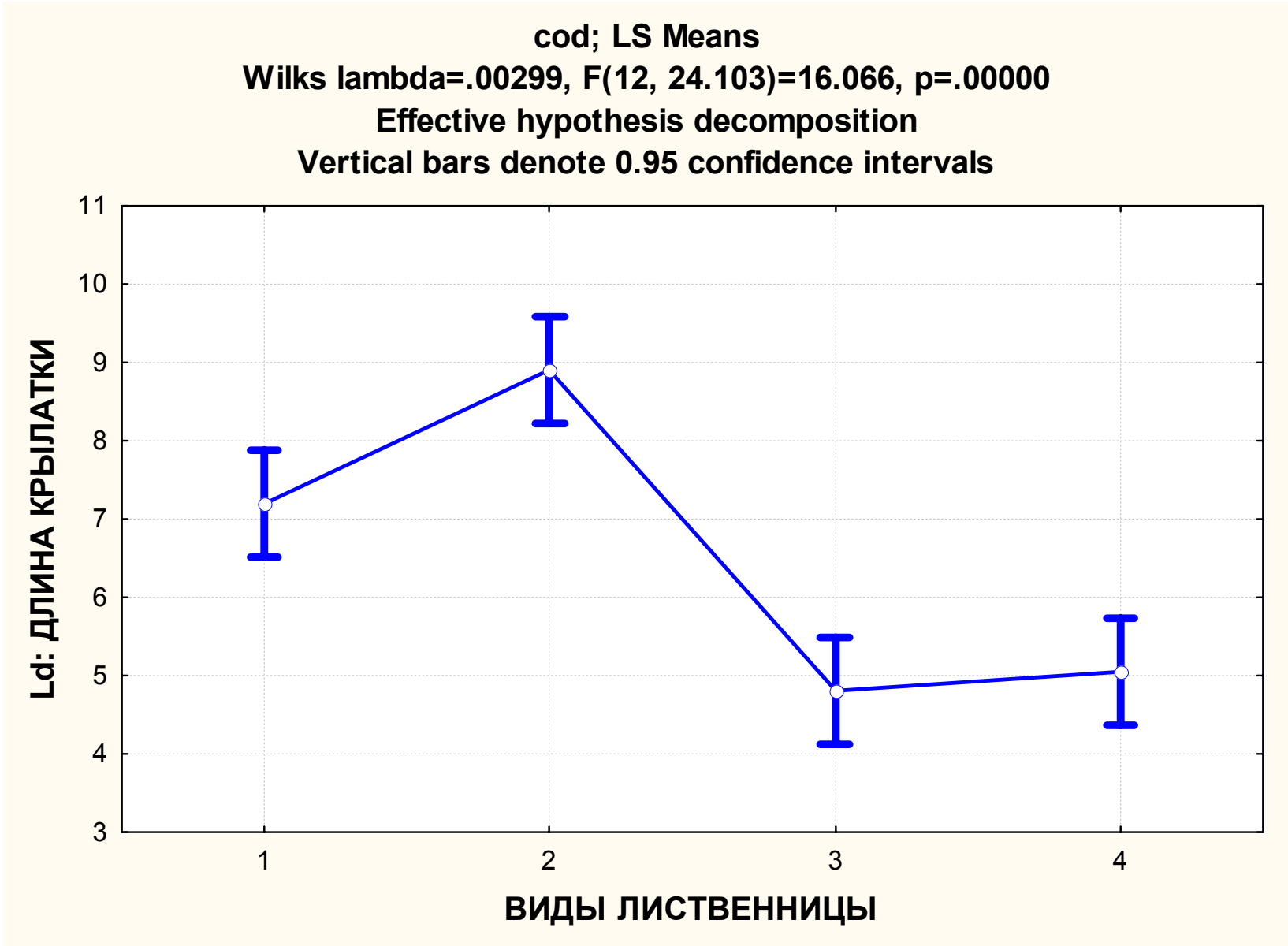

Current effect: F(4, 429)=4.9164, p=.00069
Vertical bars denote 0.95 confidence intervals
KLu: - длина верхней колосковой чешуи
3.5
3.4
3.3
3.2
3.1
3.0
2.9
2.8
2.7
2.6
2.5
0
1
2
3
4
5
6
GR
0
1

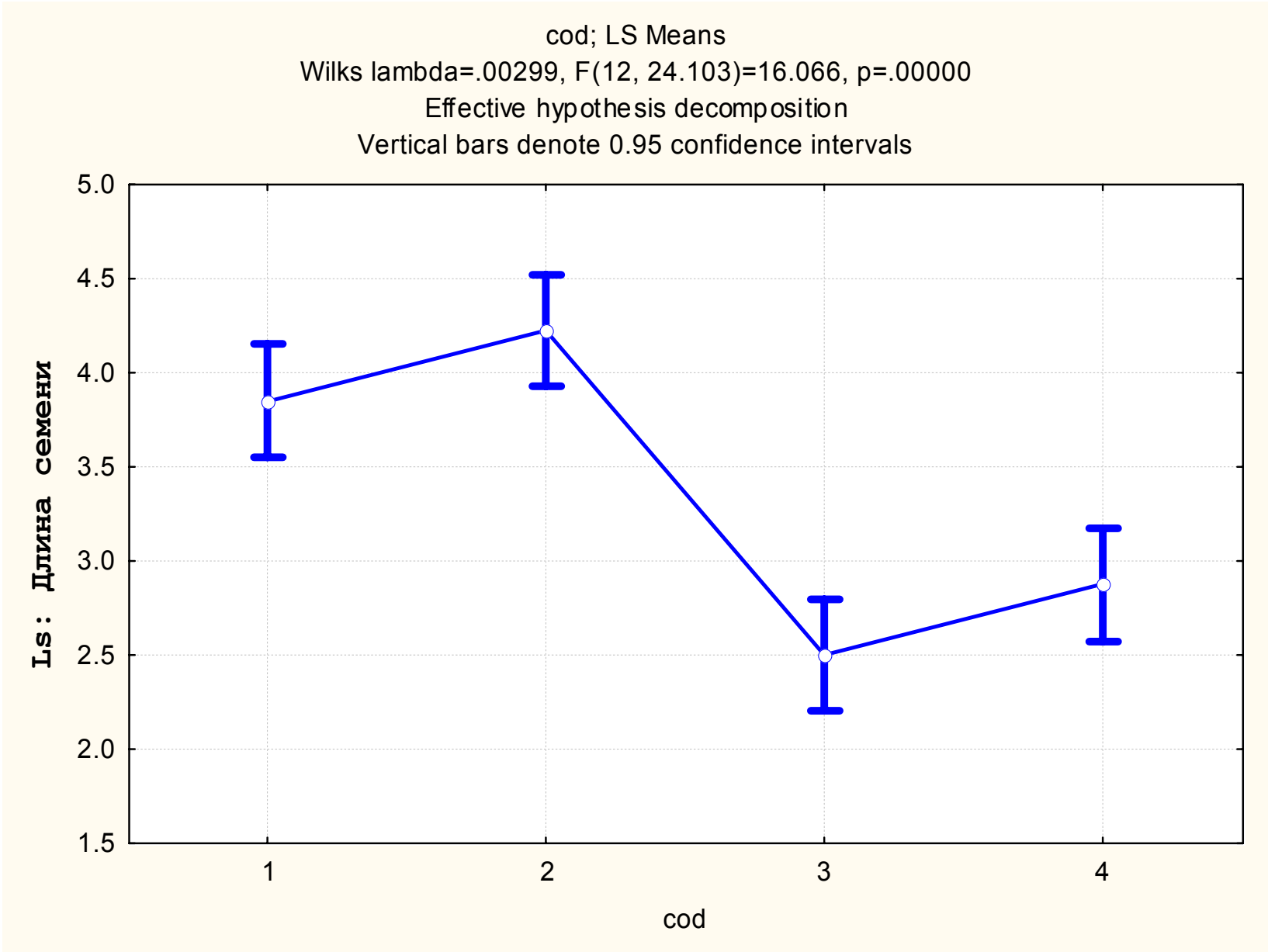
По комплексным переменным

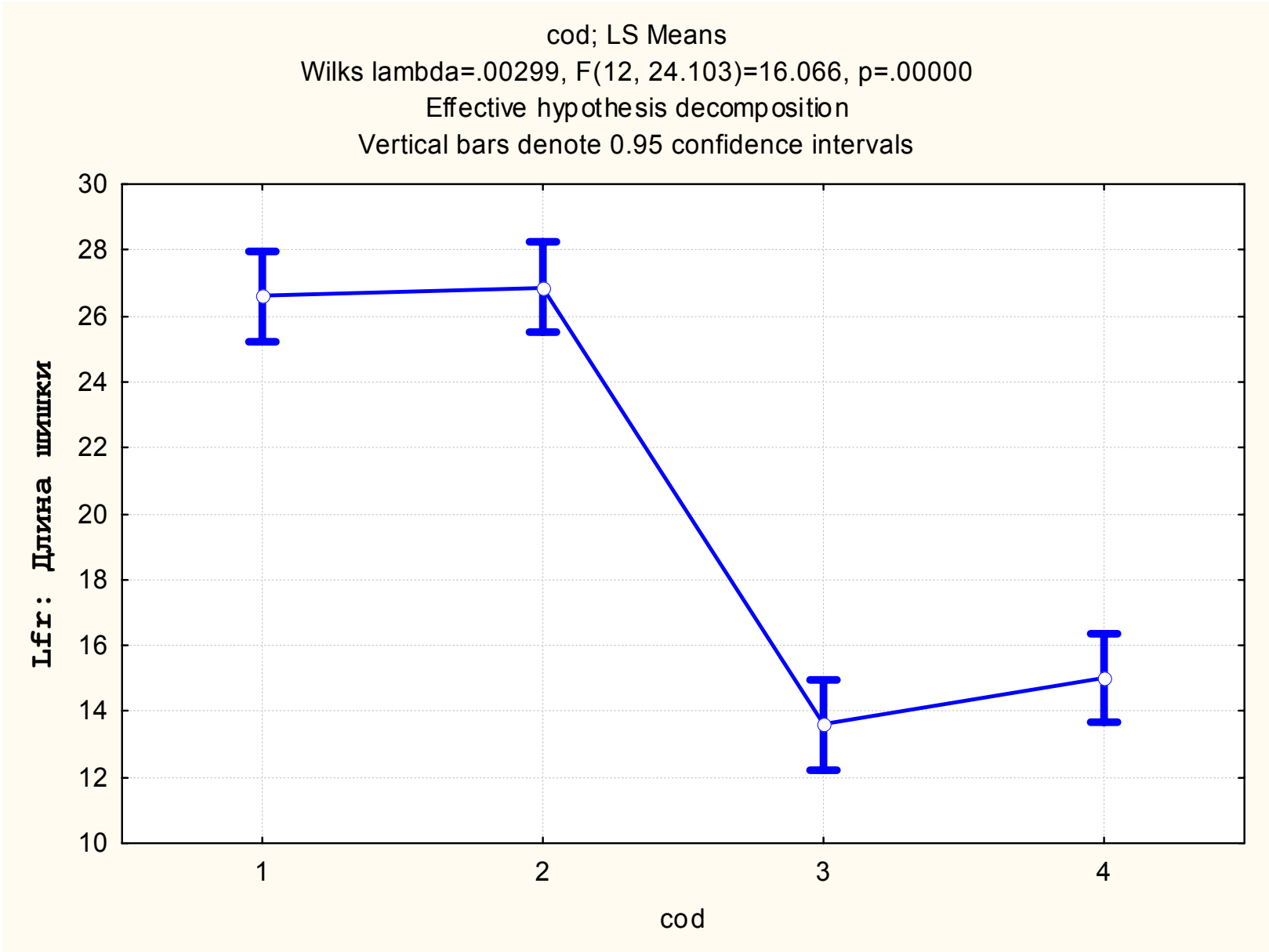


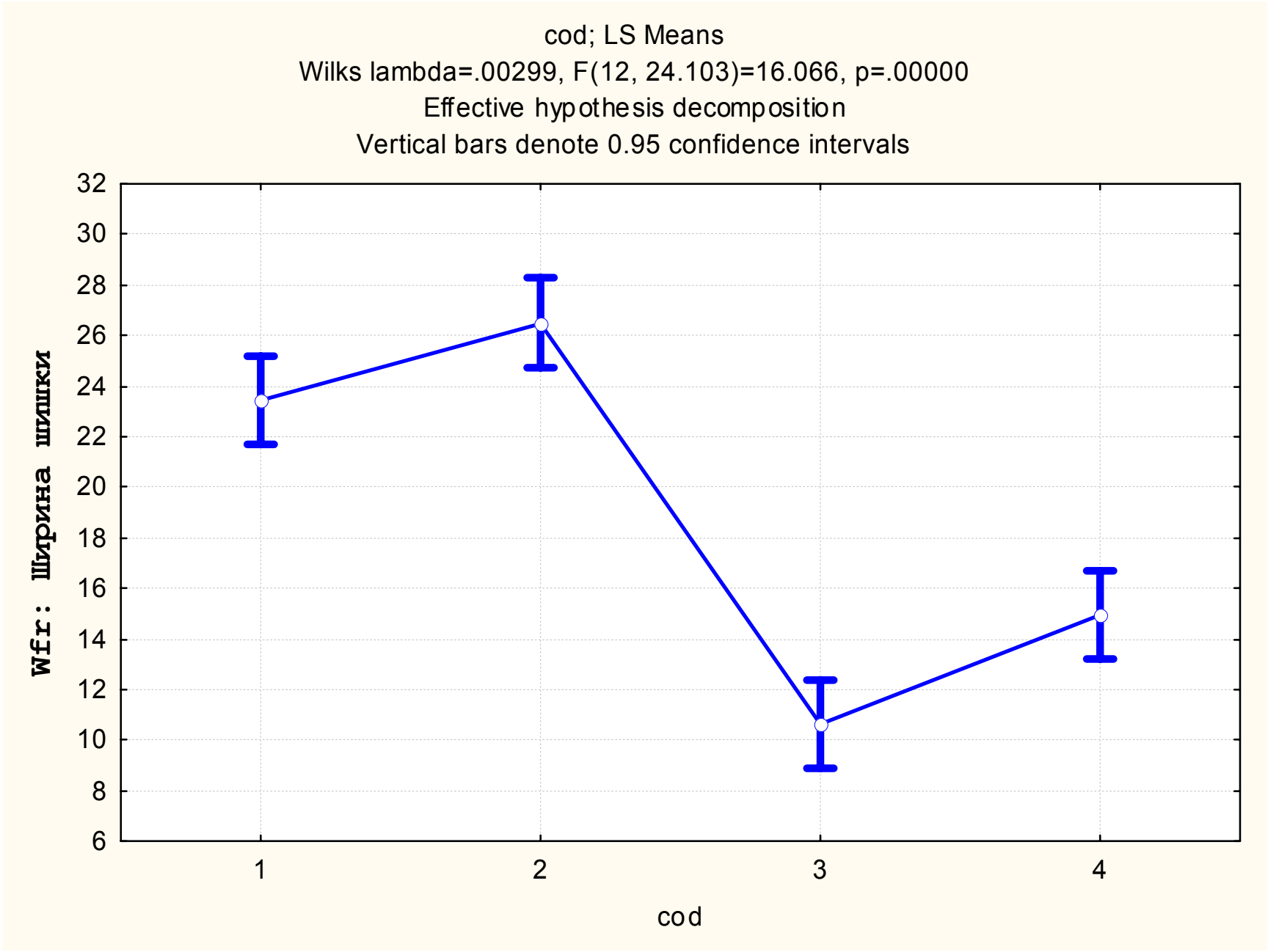

Wilks lambda=.96065, F(8, 856)=2.1696, p=.02768
1.0
0.5
0.0
-0.5
-1.0
-1.5
FACTOR2
0
1
2
3
4
5
6
GR

Изменчивость видов лиственницы в Средней Сибири


cod; LS Means
Wilks lambda=.00299, F(12, 24.103)=16.066, p=.00000
Effective hypothesis decomposition
Vertical bars denote 0.95 confidence intervals
Ld: ДЛИНА КРЫЛАТКИ
11
10
9
8
7
6
5
4
3
1
2
3
4
ВИДЫ ЛИСТВЕННИЦЫ


cod; LS Means
Wilks lambda=.00299, F(12, 24.103)=16.066, p=.00000
Effective hypothesis decomposition
Vertical bars denote 0.95 confidence intervals
Ls: Длина семени
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1
2
3
4
cod


cod; LS Means
Wilks lambda=.00299, F(12, 24.103)=16.066, p=.00000
Effective hypothesis decomposition
Vertical bars denote 0.95 confidence intervals
Lfr: Длина шишки
30
28
26
24
22
20
18
16
14
12
10
1
2
3
4
cod


cod; LS Means
Wilks lambda=.00299, F(12, 24.103)=16.066, p=.00000
Effective hypothesis decomposition
Vertical bars denote 0.95 confidence intervals
Wfr: Ширина шишки
32
30
28
26
24
22
20
18
16
14
12
10
8
6
1
2
3
4
cod

Multivariate Tests of Significance (Larix_BAR2)

	Test	Value	F	Ef-fect df	Error df	p
Intercept	Wilks	0.002327	964.7386	4	9.00000	0.00000000
cod	Wilks	0.002989	16.0655	12	24.10326	0.00000001

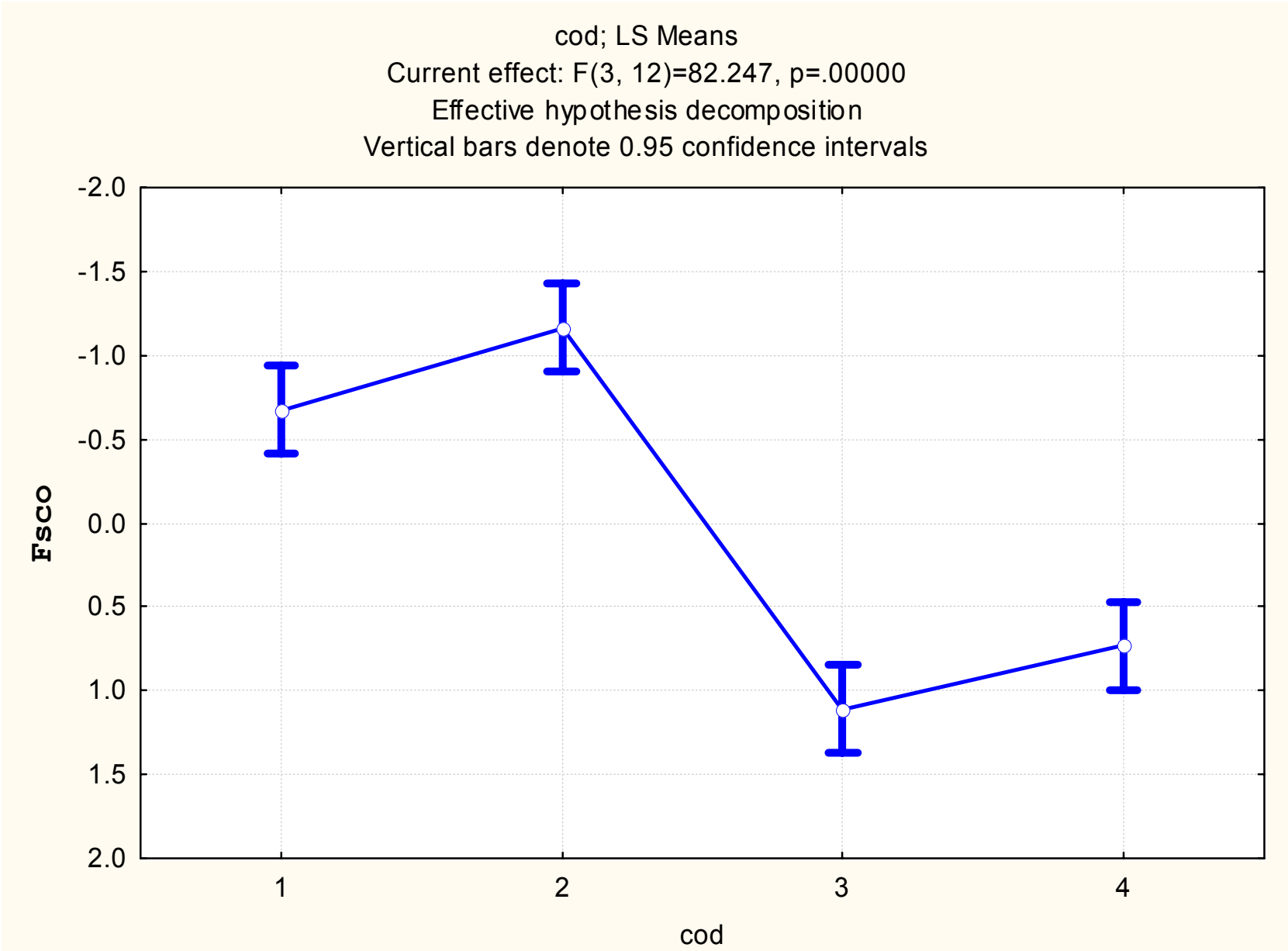
UNIVARIATE REZULTS (ЧАСТЬ)

	P_0	%%
Ld	0.00000224	90.4
Ls	0.00000341	89.7
Lfr	0.00000000	97.1
Wfr	0.00000003	95.5

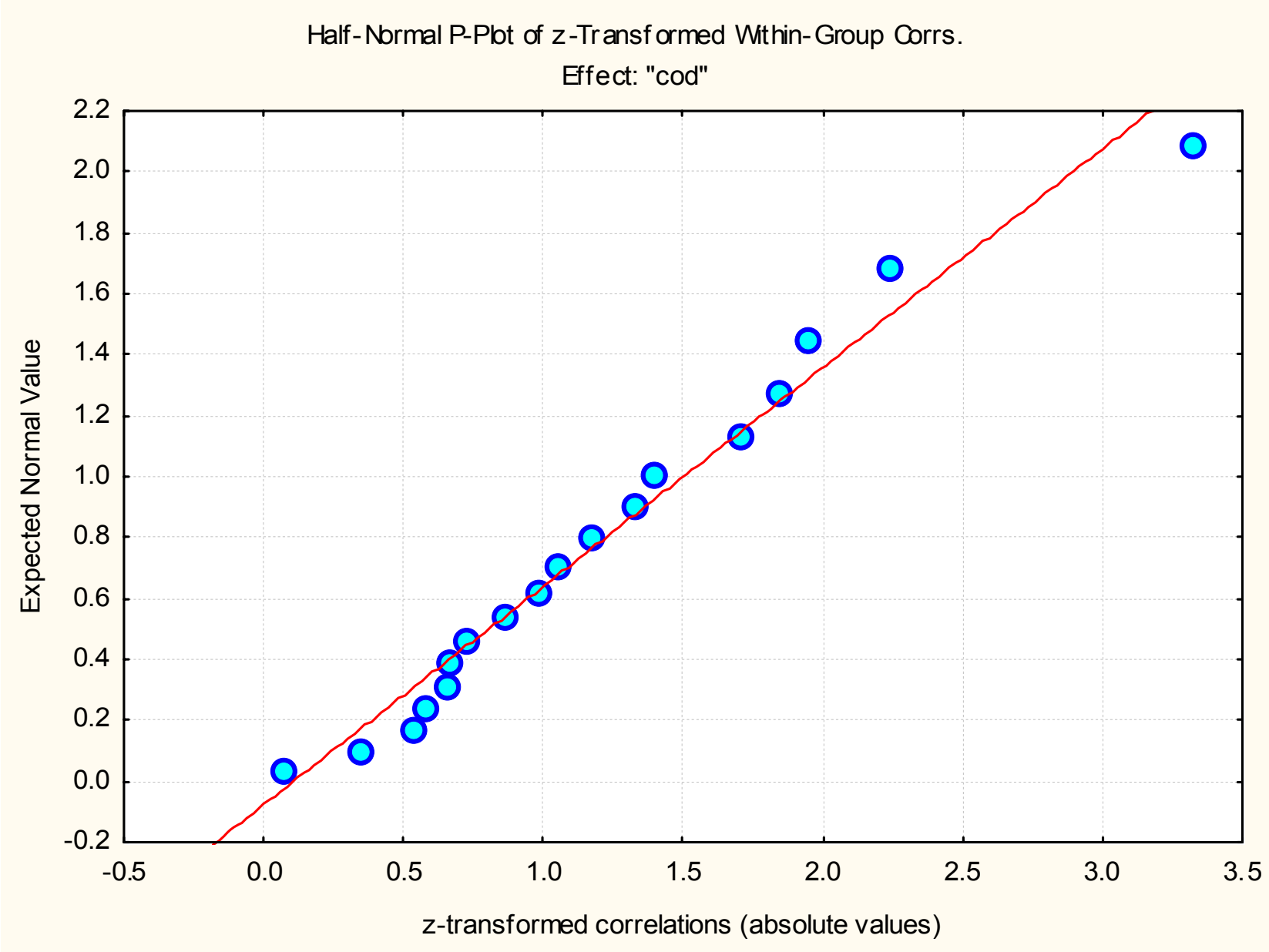
По значениям фактора - ANOVA

	d.f.	Fsco SS	Fsco MS	Fsco F	Fsco p	%%
Species	**3**	**14.3043**	**4.76811**	**82.2469**	**0.00000**	**95.40**
Error	**12**	**0.69568**	**0.05797**			**4.60**
Total	**15**	**15.0000**				

По значениям фактора



Характер распределения



- При многомерной оценке P_0 не учитываются связи (корреляции) **между** введенными в анализ зависимыми переменными!
- Существует рекомендация: не дублировать результаты, включая в анализ одновременно те переменные, которые более или менее скоррелированы друг с другом.
- Можно выбрать из группы скоррелированных признаков «признак-индикатор плеяды»

- **Более эффективно - ДО анализа**

- **1) рассмотреть структуру зависимостей и выбрать «признаки-индикаторы»,**

  или -

- **2) заменить группы скоррелированных переменных на соответствующие комплексные (интегральные) характеристики (значения факторов = factor scores) – см. В следующих лекциях = методы многомерного анализа**

| ТАК И БЫЛО СДЕЛАНО в примерах с мятликами и лиственницей|