Анализ измерений

Классификация методов

Группы методов анализа данных

Номинальные измерения: анализ частот

- Распределение (критерии согласия)
- Таблица сопряженности
- Анализ соответствий
- Логлинейный анализ таблиц сопряженности

Содержательная гипотеза: связь *X* и *Y*. *Измерения: X* и *Y номинальные переменные*

- Анализ классификации: сравнение эмпирического и теоретического (ожидаемого) распределений

Примеры: 1) Кто чаще обращается в службу знакомств: мужчины или женщины? 2) Зависит ли посещаемость занятий от дня недели? 3) Предпочитаются ли некоторые хобби чаще, чем другие?

- Анализ таблиц сопряженности: связь двух оснований классификации

Примеры: 1) Отличаются ли юноши и девушки по предпочитаемым хобби? 2) Зависит ли предпочтение одного из пяти кандидатов на выборах от пола избирателя (от его района проживания и т.п.). 3) Повлияло ли суггестивное воздействие на предпочтение одной из двух альтернатив?

Сравнение эмпирического бинарного и теоретического распределений (2-х долей): критерий согласия χ^2 (Хи-квадрат, Chi-Square) и биномиальный критерий (с. 125)

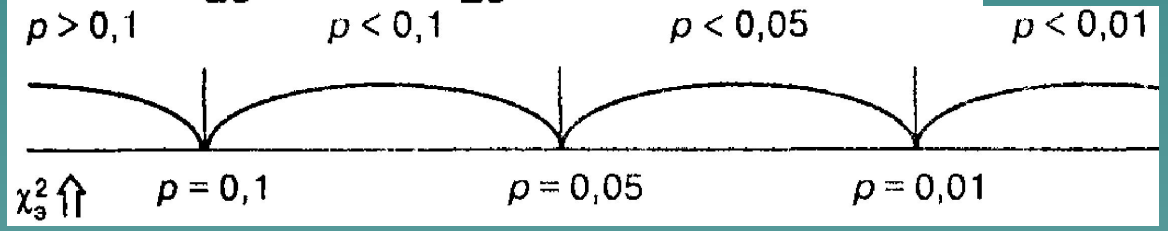
	Распределение:	
	эмпирическое	теоретическое
«За»	30	25
«Против»	20	25
Сумма:	50	50

$$\chi^2_э = \sum_{i=1}^{P} \frac{(f_э - f_т)^2}{f_т}, \quad df = (k-1)(l-1)$$

P – число ячеек с эмпирическими частотами

$(f_э)_1 = 30; (f_э)_2 = 20; (f_т)_1 = 25; (f_т)_2 = 25.$

$$\chi^2_э = \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} = 1+1 = 2, \; df = 1.$$



$p > 0,1$

А если при том же соотношении *N* = 100?

Критерий согласия Хи-квадрат: более 2-х градаций (с. 129)

С целью предсказания результатов выборов исследовалось предпочтение потенциальными избирателями пяти политических лидеров. По результатам опроса репрезентативной выборки из 120 респондентов была составлена таблица распределения их предпочтений:

Политические лидеры	Распределение предпочтений:	
	эмпирическое	теоретическое
1	21	24
2	37	24
3	29	24
4	15	24
5	18	24
Всего	120	120

Критерий согласия Хи-квадрат (SPSS)

A) Таблица частот (**Frequencies**)

var

	Observed N	Expected N	Residual
1.00	21	24.0	-3.0
2.00	37	24.0	13.0
3.00	29	24.0	5.0
4.00	15	24.0	-9.0
5.00	18	24.0	-6.0
Total	120		

Observed — эмпирические частоты, Expected — теоретические частоты.

B) Результаты статистической проверки (**Test statistics**):

Test Statistics

	Y
Chi-Square(a)	13.333
df	4
Asymp. Sig.	.010

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 24.0.

Таблицы сопряженности (с. 132)

Для каждого респондента репрезентативной выборки определены: а) пол; б) один из пяти предпочитаемых политических лидеров:

Эмпирические частоты		Y (политический лидер)					
		1	2	3	4	5	Всего:
X (пол)	муж.(1)	5	25	10	8	3	51
	жен. (2)	11	12	19	5	7	54
Всего:		16	37	29	13	10	105

Проверяется содержательная гипотеза о зависимости политических предпочтений от пола.

H_0: классификации объектов по двум основаниям являются независимыми (распределение объектов по полу не зависит от их распределения по предпочтениям политических лидеров).

Проверяем H_0 на уровне $\alpha = 0{,}05$.

$$\chi^2_э = \sum_{i=1}^{k \cdot l} \frac{(f_э - f_т)^2}{f_т}, \quad df = (k-1)(l-1)$$

Вычисление Хи-квадрат для таблиц сопряженности

Эмпирические частоты		Y (политический лидер)					
		1	2	3	4	5	Всего:
X (пол)	муж.(1)	5	25	10	8	3	51
	жен. (2)	11	12	19	5	7	54
Всего:		16	37	29	13	10	105

Теоретическая частота для ячейки *ij*: $f_{ij} = \frac{f_i \cdot f_j}{N}$

где f_i — сумма частот во всех ячейках *i*-строки; f_j — сумма частот во всех ячейках *j*-столбца; N — сумма частот всей таблицы сопряженности.

Теоретические частоты		Y (политический лидер)					
		1	2	3	4	5	Всего:
X (пол)	муж.(1)	7,77	17,97	14,09	6,31	4,86	51
	жен. (2)	8,23	19,03	14,91	6,69	5,14	54
Всего:		16	37	29	13	10	105

$$\chi_э^2 = \frac{(5-7,77)^2}{7,77} + \ldots = 11,84; df = (2-1)(5-1) = 4. \Rightarrow p < 0,05.$$

Вывод: обнаружена статистически значимая связь политических предпочтений и пола (p < 0,05)

Таблицы сопряженности 2x2 (с. 135)

1) *по двум различным* дихотомическим основаниям — случай независимых выборок;
2) *по одному и тому* же дихотомическому основанию дважды (например, до и после воздействия) — случай зависимых выборок.

ПРИМЕРЫ

1. Случай независимых выборок. Две группы больных известной численности получали курс лечения разными методами. Подсчитывалось число рецидивов заболевания в той и другой группе. Одна переменная — «метод лечения» (1-й, 2-й), другая — «рецидив» (есть, нет).
2. Случай зависимых выборок. Подсчитывалось число тех, кто «за», и тех, кто «против» смертной казни: до и после убедительной лекции о введении моратория на смертную казнь. Одна переменная — «до лекции» («за», «против»), другая переменная — «после лекции» («за», «против»).

Для независимых выборок применяется критерий χ^2-Пирсона, а для зависимых более адекватным является метод Мак-Нимара.

Таблица 2x2: независимые выборки (с. 136)

Предположим, для изучения влияния 2-х условий запоминания материала 100 испытуемых были случайным образом разделены на две группы: по 50 человек для каждого из условий. После обучения количество усвоивших этот материал в первой группе составило 24 человека, а во второй — 34 человека. Можно ли утверждать, что различия в условиях влияют на результативность обучения?

ВАЖНО: ДАННЫЕ ПРЕДСТАВЛЯЮТ СОБОЙ ДВЕ ПЕРЕМЕННЫЕ!

	Усвоение материала		Всего:
	есть	нет	
Условие 1	24	26	50
Условие 2	34	16	50
Всего:	58	42	100

Критерий Хи-квадрат **с поправкой на непрерывность**:

$$\chi^2_э = \sum_{i=1}^{4} \frac{(|f_э - f_т| - 0{,}5)^2}{f_т}, \; df = 1.$$

Допускается 1-сторонняя альтернатива!

Теоретические частоты:

$$f_{11} = \frac{50 \cdot 58}{100} = 29, \; f_{12} = \frac{50 \cdot 42}{100} = 21, \; f_{21} = \frac{50 \cdot 58}{100} = 29, \; f_{22} = \frac{50 \cdot 42}{100} = 21.$$

$$\chi_э^2 = 3{,}325; df = 1.$$

Альтернатива: 2-сторонняя или 1-сторонняя?

Таблицы 2x2: повторные измерения бинарной переменной (с. 139)

Исследовалось влияние убедительной лекции о введении моратория на смертную казнь. Число респондентов $N = 60$. Подсчитывалось число тех, кто «за», и тех, кто «против» смертной казни до и после лекции. Одна переменная — «до лекции» («за», «против»), другая — «после лекции» («за», «против»).

		До:		Всего:
		«За»	«Против»	
После:	«За»	$a = 16$	$b = 10$	26
	«Против»	$c = 26$	$d = 8$	34
Всего:		42	18	60

Критерий Хи-квадрат не применим! $\chi^2_э = 0{,}93,\ df = 1,\ p > 0{,}1$

Критерий Мак-Нимара:

$$z_э = \frac{c-b}{\sqrt{c+b}} \text{ или } z_э = \frac{a-d}{\sqrt{a+d}}$$

$$z_э = \frac{c-b}{\sqrt{c+b}} = \frac{26-10}{\sqrt{26+10}} = 2{,}67$$

p - ?

Сравнительный анализ

- Методы сравнения двух выборок
- Однофакторный ANOVA и непараметрические аналоги
- Многофакторный ANOVA
- Многомерный ANOVA
- Дискриминантный анализ
- ANOVA с повторными измерениями

Классификация методов сравнения (с. 113)

Количество выборок (градаций X)		Две выборки		Больше двух выборок	
Зависимость выборок		Независимые	Зависимые	Независимые	Зависимые
Признак Y	метрический	Параметрические методы сравнения			
		t-Стьюдента, для независимых выборок	t-Стьюдента, для зависимых выборок	ANOVA	ANOVA, с повторными измерениями
	ранговый	Непараметрические методы сравнения			
		U-Манна-Уитни, критерий серий	T-Вилкоксона, критерий знаков	H-Краскала-Уоллеса	χ^2-Фридмана

Если Y – *метрическая переменная* (распределение приблизительно нормальное), то применяются *методы сравнения средних*.

Если Y – *порядковая переменная* (выбросы, асимметрия распределения…), *или* $N < 20\text{-}25$, то применяются *ранговые методы (критерии) сравнения*, предполагающие предварительное ранжирование Y.

Разработал Р.Фишер (1920-е гг.) – для анализа экспериментальных данных

Основные понятия:

Фактор (*X* - независимая переменная) – группирующая, номинальная, характеризуется уровнями (градациями).

Уровень = группа (выборка).

Зависимая переменная – (*Y*) – метрическая.

Т.о. каждому уровню фактора соответствует среднее значение зависимой переменной.

Межгрупповые факторы – уровням соответствуют независимые выборки.

Внутригрупповые факторы – уровням соответствуют зависимые выборки.

Фиксированные и *случайные* факторы.

Ковариата – метрическая независимая переменная, «включаемая» в анализ наряду с фактором.

Принципиальная идея ANOVA

В дисперсии зависимой переменной выделяется две составляющие:

межгрупповая (D_f) – влияние фактора

и внутригрупповая (D_e) – остальные причины.

$$D = D_f + D_e$$

Чем сильнее различаются групповые средние, тем больше D_f.
Чем выше изменчивость внутри каждой группы, тем выше D_e.
Статистическая значимость определяется соотношением D_f / D_e.
Величина эффекта: $R^2 = D_f / D$

Виды ANOVA и их специфические проблемы

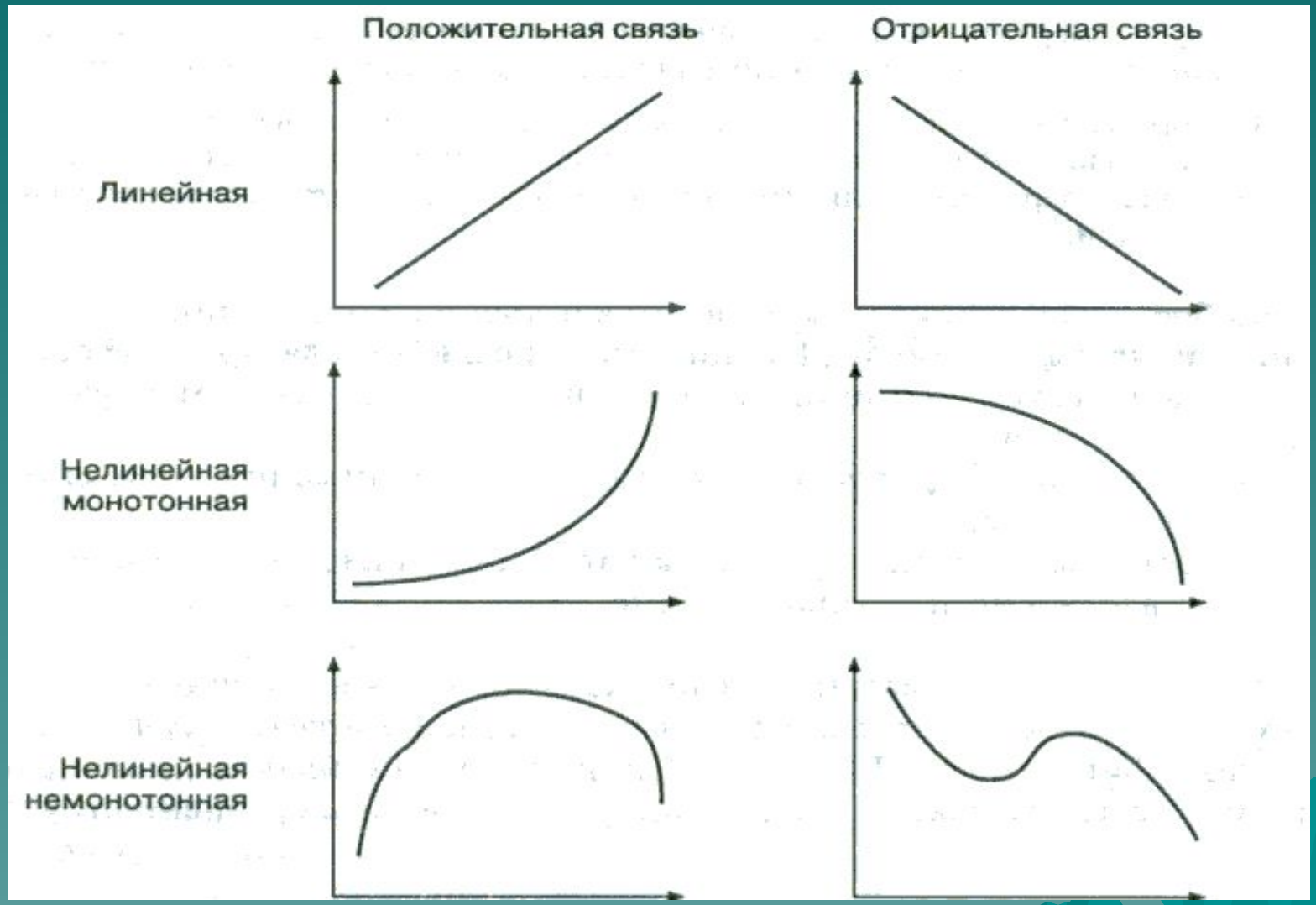
1. Однофакторный ANOVA: множественные сравнения средних.
2. Многофакторный ANOVA: главные эффекты и взаимодействия факторов.
3. Многомерный ANOVA (MANOVA): применение многомерных критериев.
4. ANOVA с повторными измерениями: межгрупповые и внутригрупповые эффекты.

2 – 4: Общие Линейные Модели - ОЛМ (General Linear Models - GLM)

Коэффициент корреляции

r - мера вероятностной связи двух количественных переменных

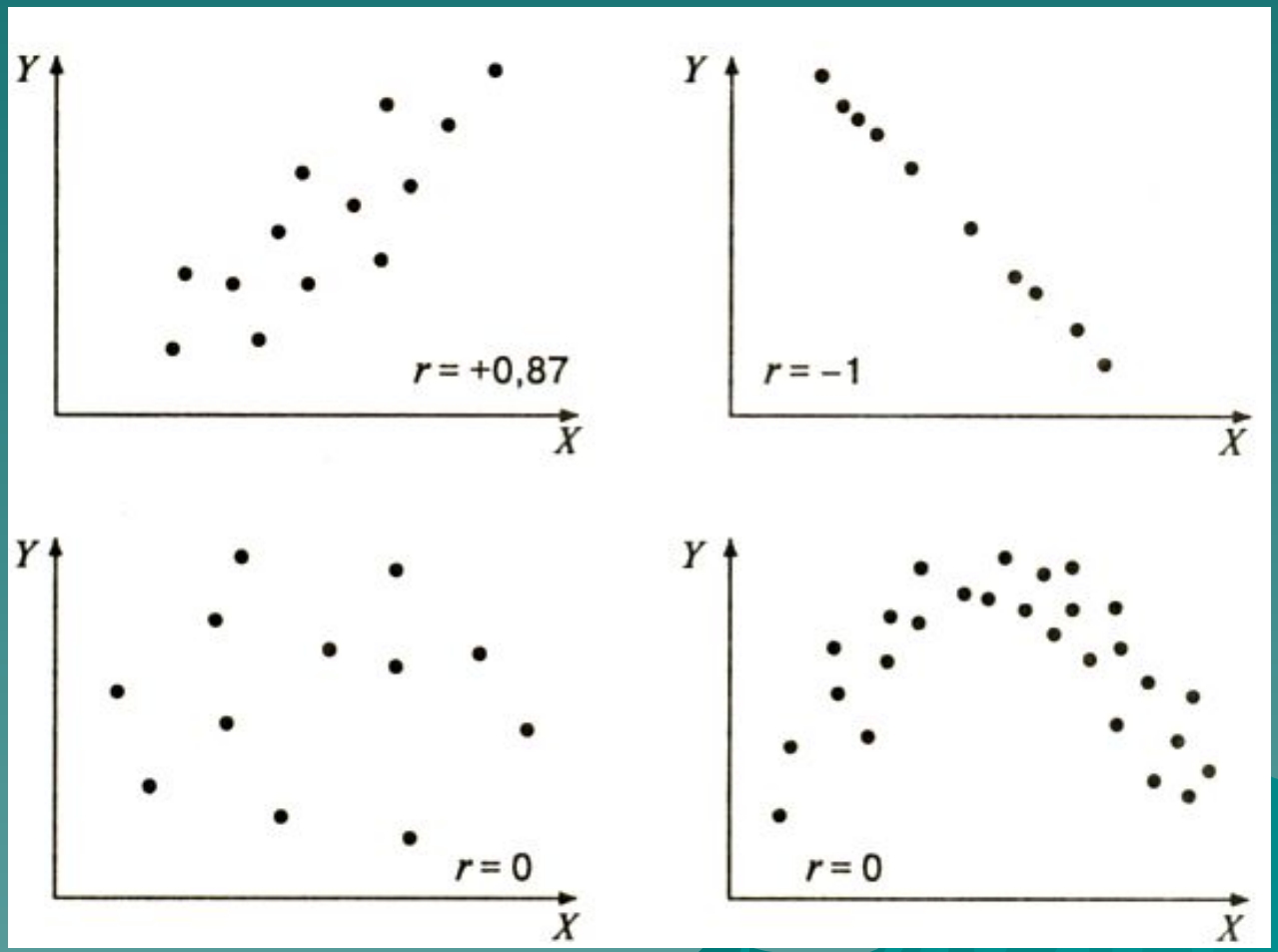
Связи: функциональные …



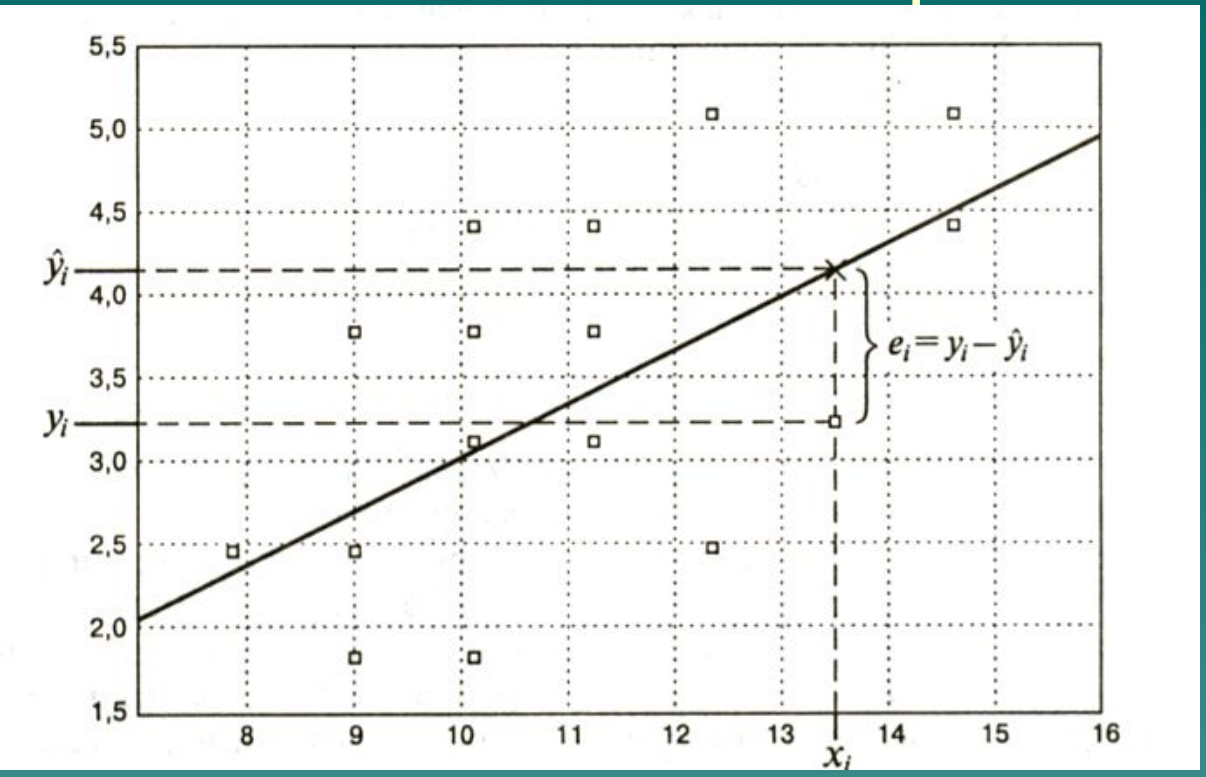
…и статистические

Коэффициент корреляции ***r*** это количественная мера силы (абсолютное значение) и направления (знак) вероятностной взаимосвязи двух переменных.

$-1 \leq r \leq +1$



Регрессия



Уравнение регрессии:

$$\hat{y}_i = bx_i + a$$

Коэффициент регрессии:

$$b = r_{xy}\frac{\sigma_y}{\sigma_x}$$

Свободный член:

$$a = M_y - bM_x$$

y_i — истинное i-значение Y,

$\hat{y}_i$ — *оценка i-значения Y по значению* x_i *при помощи линии (уравнения) регрессии*,

$e_i = y_i - \hat{y}_i$ – *ошибка оценки*

Линия регрессии (прямая) аппроксимирует точки методом наименьших квадратов: $\sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2 = \min$

Коэффициент детерминации

Дисперсия оценок зависимой переменной *Y* $\sigma^2_{\hat{y}}$ – часть её дисперсии σ^2_y, обусловленная влиянием независимой переменной *X*: $0 \le \sigma^2_{\hat{y}} \le \sigma^2_y$

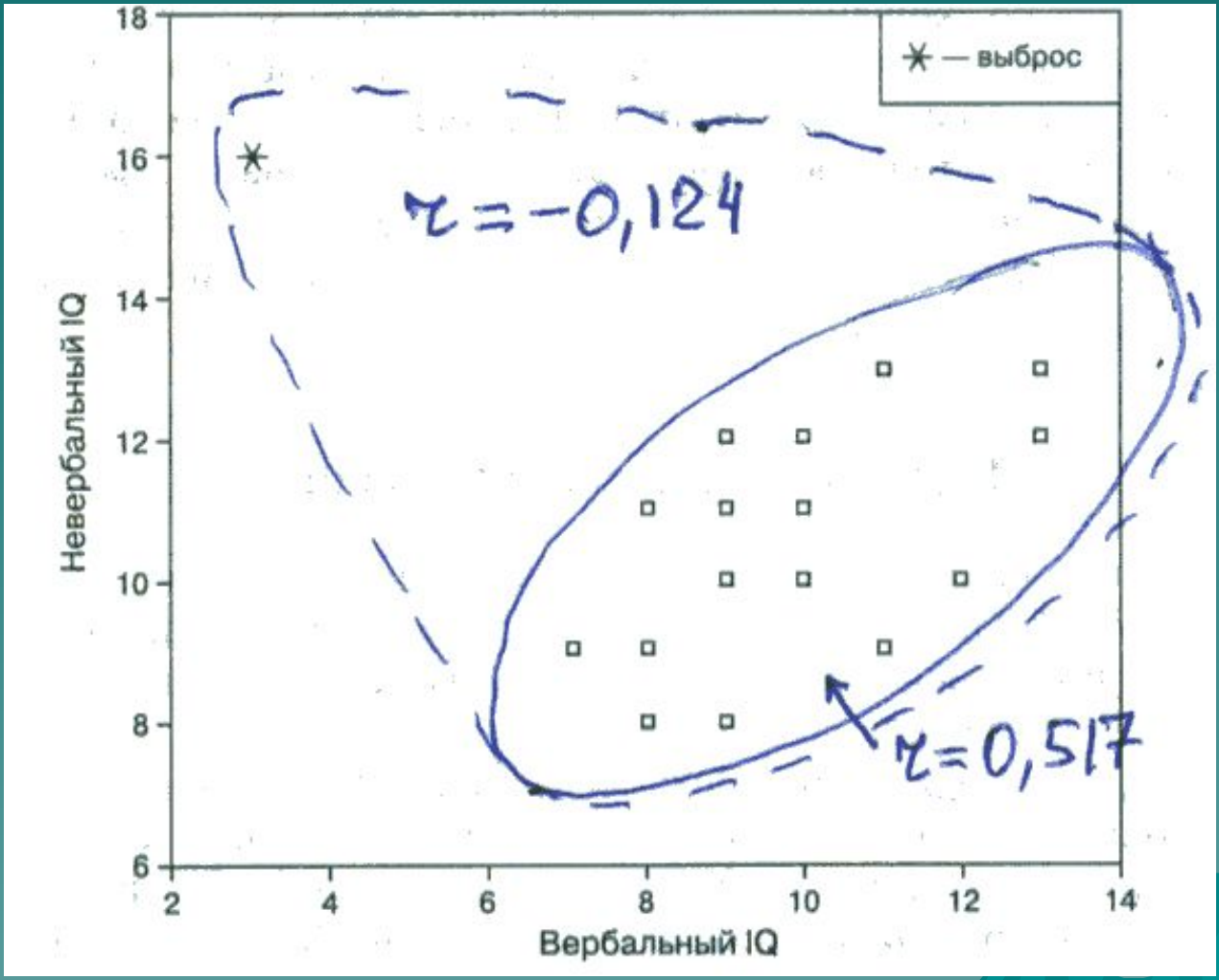
$$\sigma^2_{\hat{y}_i} = \sigma^2_{bx_i+a} = \sigma^2_{bx_i} = b^2\sigma^2_{x_i} = r_{xy}{}^2 \frac{\sigma^2_{y_i}}{\sigma^2_{x_i}} \sigma^2_{x_i} = r_{xy}{}^2 \sigma_{y_i}{}^2$$

$$r_{xy}{}^2 = \frac{\sigma_{\hat{y}_i}{}^2}{\sigma_{y_i}{}^2}$$

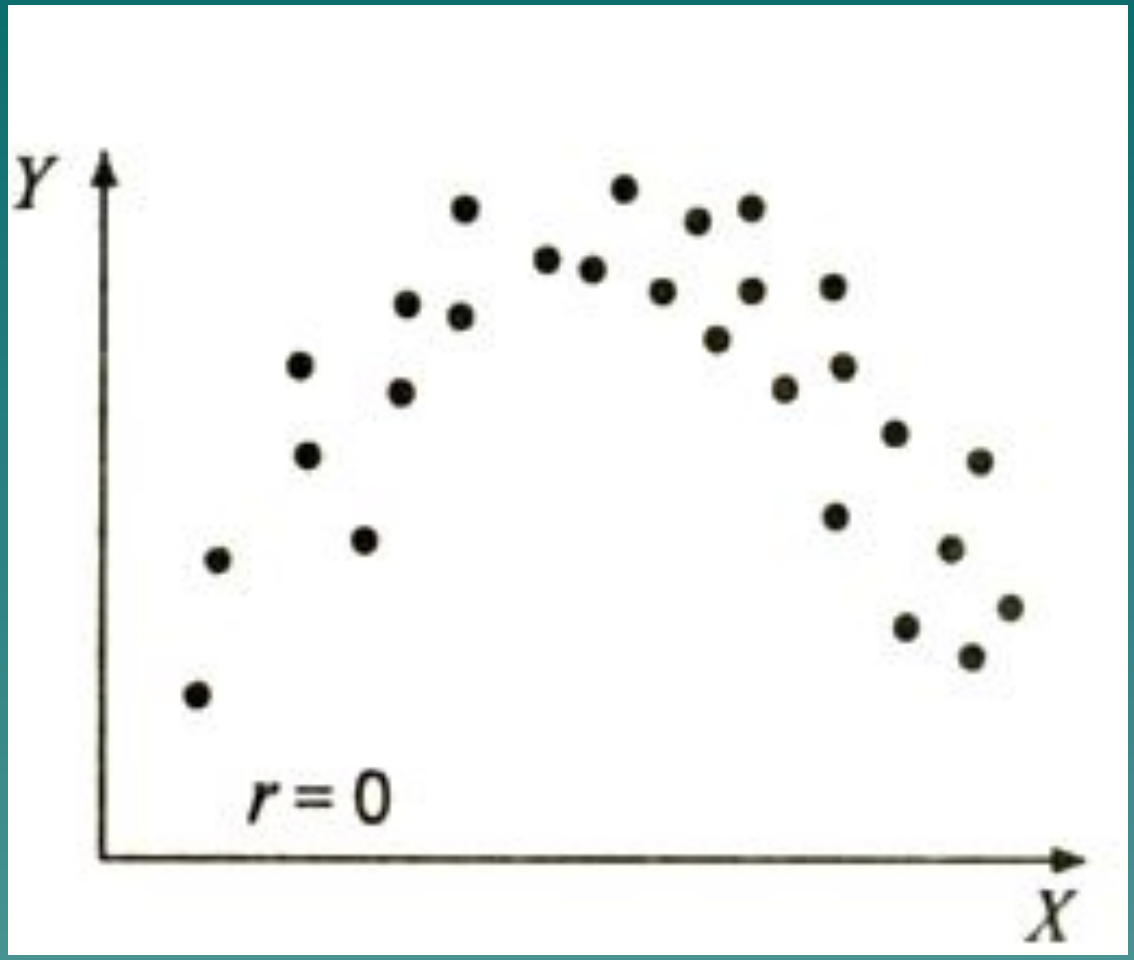
- коэффициент детерминации, доля дисперсии переменной *Y* (от 1), «объясняемая» влиянием переменной *X*.

Величина корреляции и сила связи

1) выбросы и асимметрии распределений

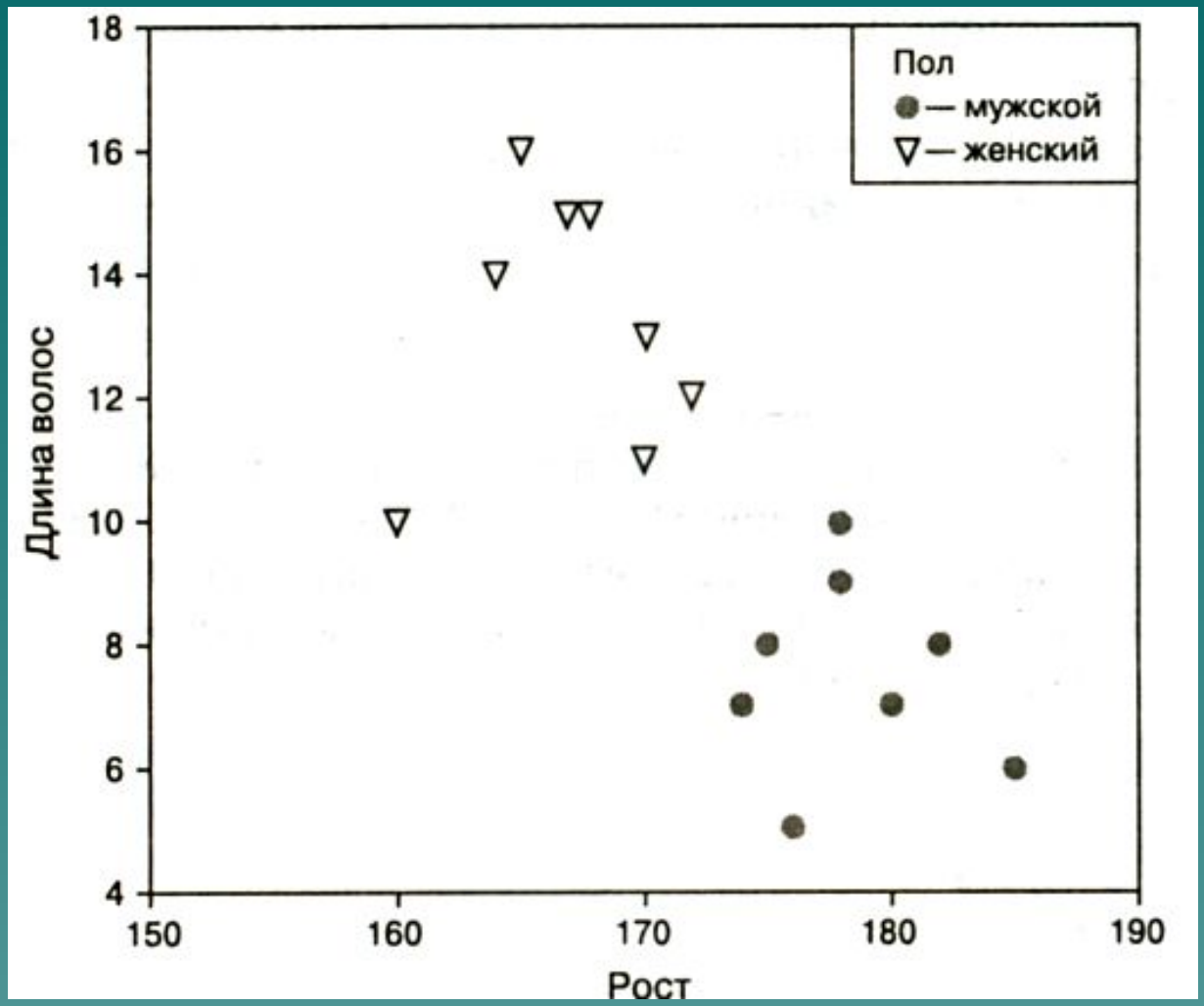


2) Нелинейные связи



$$\hat{y}_i = a + b_1 x_i + b_2 x_i^2 \Longrightarrow \quad R^2 \approx 0,7$$

3) Влияние «третьей» переменной



$$r \approx -0,7(!)$$

Частная корреляция

Корреляция IQ (x) и длины стопы (y) $r_{xy} = 0,42$

но корреляция IQ с возрастом (z) $r_{xz} = 0,6$
а корреляция возраста и длины стопы $r_{yz} = 0,7$

$$r_{xy-z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

$$r_{xy-z} = \frac{0.42 - 0.7 \times 0.6}{\sqrt{(1 - 0.7^2)(1 - 0.6^2)}} = 0$$

Ранговые корреляции

Вычисляются после замены исходных

значений рангами.

r-Спирмена, аналог r-Пирсона, основан на разности рангов

т-Кендалла, вероятностный, основан на подсчете совпадений и инверсий в парах наблюдений.

Оцените величину корреляции без вычислений

1)

№	X	Y
1	1	1
2	0	0
3	0	0
4	1	1
5	1	1
6	0	0

2)

№	X	Y
1	1	0
2	0	1
3	0	1
4	1	0
5	1	0
6	0	1

3)

№	X	Y
1	3	16
2	2	14
3	4	18
4	1	12
5	6	22
6	5	20

4)

№	X	Y
1	36	100
2	34	103
3	38	97
4	32	106
5	42	91
6	40	94

5)

№	X	Y
1	8	124
2	10	500
3	6	88
4	2	58
5	4	66
6	0	52

6)

№	X	Y
1	27	24
2	35	7
3	23	98
4	16	123
5	17	99
6	12	158

1 и 2 – чему равен ϕ-сопряженности? r-Пирсона? ранговая корреляция?
3 - 6 – чему равен r-Пирсона? ранговая корреляция?
Варианты ответов:
а) = 1; б) = -1; в) отрицательный, но > -1; г) положительный, но < 1.

Последовательность интерпретации корреляций

1. Статистическая значимость (p-уровень).
2. Знак (направление).
3. Величина (по *r*-квадрат).

Числовые показатели: *r* = …; *N* = …; *p* = … .

ПРИМЕР. Для проверки гипотезы … применялась корреляция Пирсона. Обнаружена статистически достоверная отрицательная корреляция показателей тревожности и креативности (r = -0,435; N = 32; p = 0,035): чем выше тревожность, тем ниже креативность.

Корреляционная матрица

Корреляции

		тест1	тест2	тест3	тест4	тест5
тест1	Корреляция Пирсона	1	,436**	-,051	-,134	-0,201**
	Знч.(2-сторон)		,000	,617	,185	,047
	N	100	100	100	100	100
тест2	Корреляция Пирсона	,436**	1	-,196	-,153	,015
	Знч.(2-сторон)	,000		,050	,127	,886
	N	100	100	100	100	100
тест3	Корреляция Пирсона	-,051	-,196	1	,441**	,483**
	Знч.(2-сторон)	,617	,050		,000	,000
	N	100	100	100	100	100
тест4	Корреляция Пирсона	-,134	-,153	,441**	1	,475**
	Знч.(2-сторон)	,185	,127	,000		,000
	N	100	100	100	100	100
тест5	Корреляция Пирсона	-0,201*	,015	,483**	,475**	1
	Знч.(2-сторон)	,047	,886	,000	,000	
	N	100	100	100	100	100

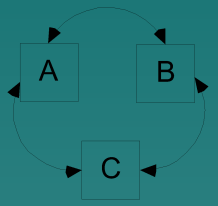
**. Корреляция значима на уровне 0.01 (2-сторон.).
*. Корреляция значима на уровне 0.05 (2-сторон.).

Поправка Benjamini & Hochberg (1995) для семейства *n* гипотез

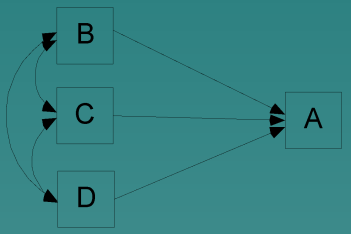
1) Упорядочиваем все p от min до max (*i* – текущий номер p в ряду);
2) Для каждого *i* вычисляем: $p*n/i = p_{корр.}$;
3) Если $p_{корр.} \leq \alpha$ – результат статистически достоверен!

Корреляционный анализ

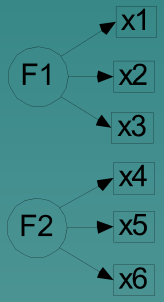
- Корреляционные матрицы, плеяды, частная корреляция и анализ криволинейности



- Множественный регрессионный
- анализ



- Факторный анализ



- Структурное моделирование