ОСНОВЫ БИОСТАТИСТИКИ

Александр Владимирович Рубанович

зав. лаб. экологической генетики ИОГен РАН

rubanovich@vigg.ru

тел. (499) 132-8958

Темы для обсуждения

- Оценка ассоциаций «генотип-фенотип» и их значимости
- Факторы, влияющие на значимость оценок
- Объединение выборок и метаисследования
- Учет множественности сравнений

Выявление ассоциаций «генотип-фенотип»: минимальный набор действий

- Фено
  (на
  н»)

Кроме этого в обоих случаях можно строить различные регрессионные модели: Зависимая переменная – признак (фенотип), независимыми переменные – генотипы. Например так: A/A - 0, A/T - 1, T/T - 2

Выч
ра

- Фенотип - количественный признак
  (например: вес, содержание кальция, частота аберраций)

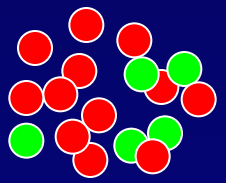
Вычисляем средние значения признака для разных генотипов; значимость по критерию Манна-Уитни

OR – количественная мера предрасположенности (Odd Ratio)

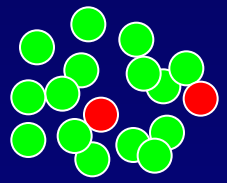
OR – непременный атрибут «**case-control association study**» (выявление «генов предрасположенности» к заболеванию путем сопоставлений частот генотипов у больных и здоровых)

OR показывает во сколько раз повышена вероятность заболеть для носителя «плохого» генотипа

Группа больных



Контроль (здоровые)



● - генотип, указывающий на предрасположенность к заболевания

$P_{больные} >> P_{контроль}$

$$OR = \frac{P_{больные}(1 - P_{контроль})}{P_{контроль}(1 - P_{больные})}$$

OR>1 – генотип связан с болезнью
OR=1 – нет связи между генотипом и болезнью
OR<1 – протективный генотип

Soft для вычисления OR и проведения матаисследований

WinPepi Portal (2010) - computer programs for epidemiologists

- 

Free!

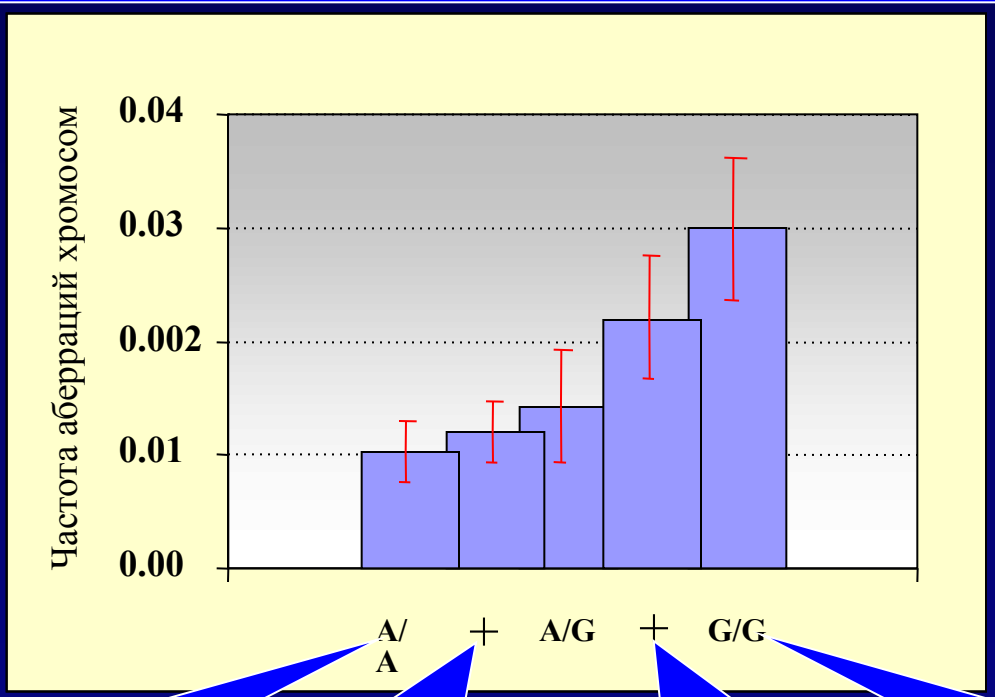
- 

30 дней

Статистический анализ сопряженности генотипов и количественных признаков

- Самое простое и необходимое: вычисление средних значений признака для носителей различных генотипов. Далее сравнение по непараметрическому тесту

Обычно стараются рассмотреть две группы



Гомозигота по мажорному аллелю

сивная дель

Домина моде

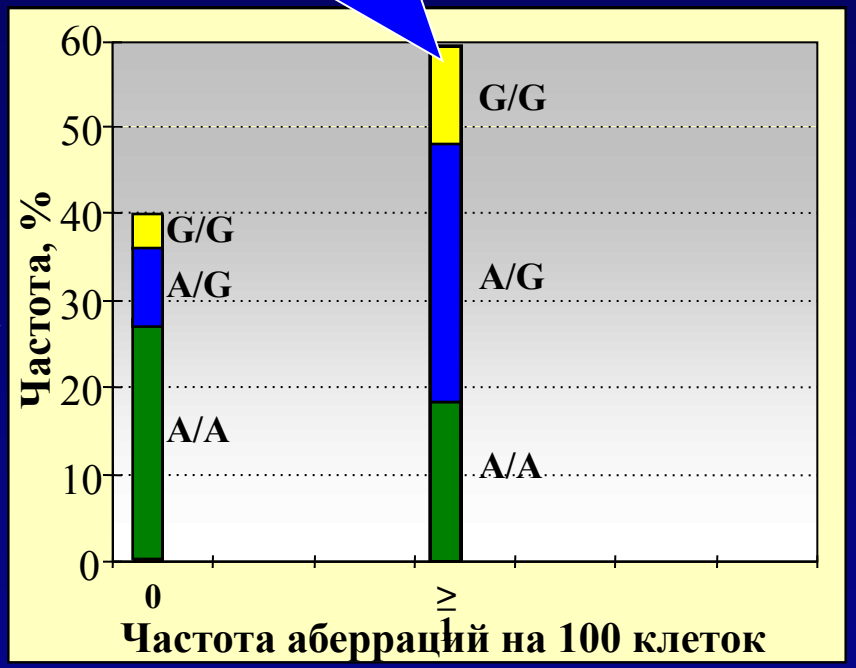
Гомозигота по минорному аллелю

Статистический анализ сопряженности генотипов и количественных признаков

- Самое простое и необходимое: вычисление средних значений
  Далее сра
  (не по Ст
  отипов.

Далее вычисляется OR
и значимость по точному критерию Фишера.
В данном примере риск возникновения аберраций у носителей минорного аллеля G равен OR=2,1 и p=0,015

- Сравнение частот генотипов для
  низким (или высоким) значением признака



Статистический анализ сопряженности генотипов и количественных признаков

- Самое простое и необходимое: вычисление средних значений признака для носителей различных генотипов. Далее сравнение по непараметрическому тесту (не по Стьюденту!)
- С… …ким (или вы…
- Логистическая и пуассоновская регрессии

Зависимая переменная – признак (p), независимыми переменные – генотипы (x_i). Например так: A/A - 0, A/T - 1, T/T - 2

$$p = \frac{1}{1 + e^{a_1 x_1 + \ldots + a_n x_n}}$$

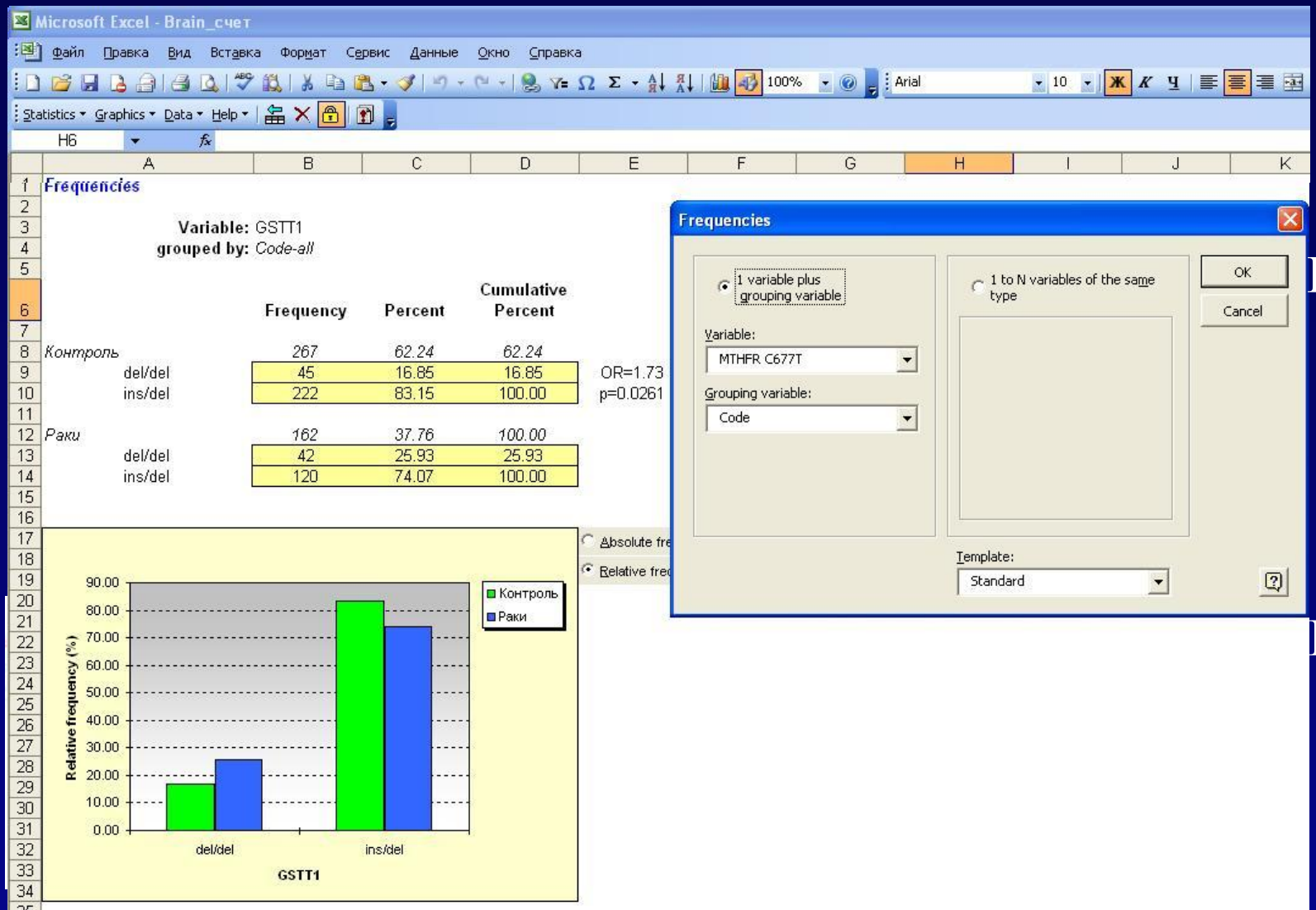
p – частота аберраций
x_i – генотип i-го локуса
a_i – коэф. регрессии

Для логистической регрессии $a_i = ln(OR_i)$

$$p = e^{a_0 + a_1 x_1 + \ldots + a_n x_n}$$

Soft для работы с генотипами и гаплотипами

- **WinStat for Excel**



- Free!

- Free!

Темы для обсуждения

- Оценка ассоциаций «генотип-фенотип» и их значимости
- Факторы, влияющие на значимость оценок
- Объединение выборок и метаисследования
- Учет множественности сравнений

Чуть-чуть об ошибках статистических тестов

Нулевая **различи** **ии** **ости**

Традиционно биолог ориентирован на контроль
ошибки I рода (через уровень значимости),
т.е. на гарантии отсутствия ложных открытий,

Ошибка I рода (α)

Вероятность отвергнуть правильную нулевую гипотезу =
Вероятность обнаружить различия там, где их нет = Вероя
совершить фальшивое открытие

Ошибка II рода (β)

Вероятность принять неправильную нулевую гипотезу =
Вероятность не обнаружить существующие различия =
Вероятность упустить открытие

Мощнос

Вероятн

Вероятн

… и при этом мало заботится о возможности
упустить открытие (ошибка II рода)

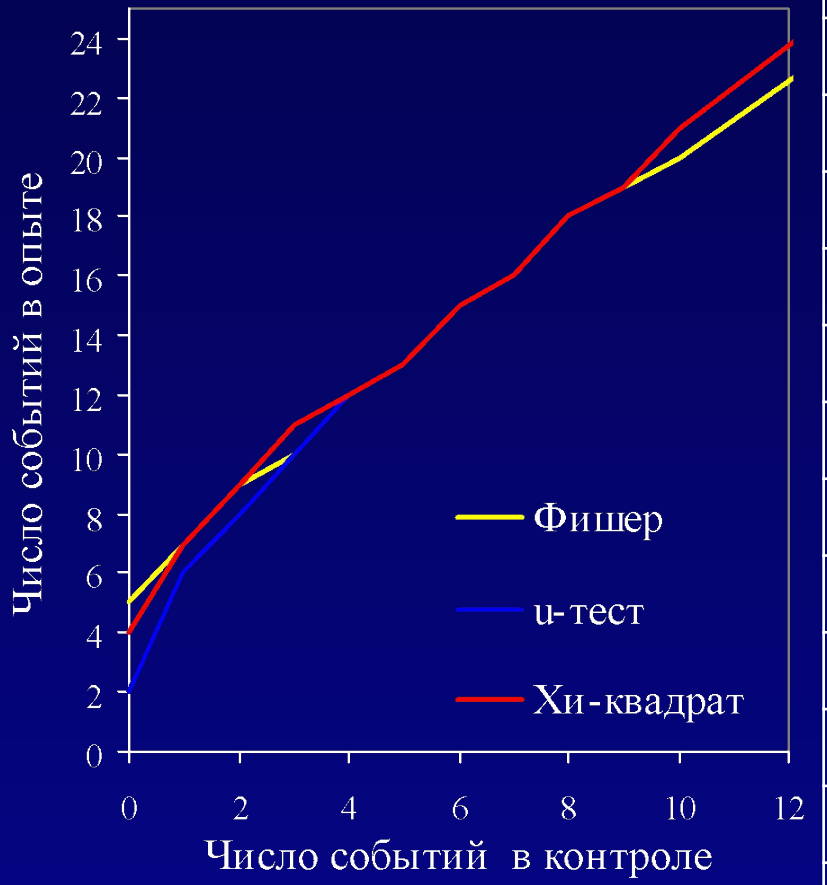
От чего зависят ошибки статистических тестов?

- От размаха реально существующих отличий и разброса данных
- От объемов выборок
- Ошибка I рода (вероятность фальшивого открытия) слабо зависит от объемов выборок, если они сравнимы по величине

Крайний случай:
«критерий» св. Фомы Неверующего (0033)
Ошибка I рода = 0 ⇔ Ошибка II рода = 1

Сравнение частот при уровне значимости 0.05

Объемы выборок в опыте и контроле одинаковы



Число событий в контроле	*Минимальное число событий в опыте при значимом отличии от контроля*		
	Стьюдент	χ^2	*Фишер*
0	2	4	5
1	6	7	7
2		9	
6	15	15	15
7	16	16	16
8	18	18	18
9	19	19	19
10	21	21	20
20	35	35	33
30	47	47	46

больше 5
независимо от объемов выборок
(100 или 1000)

Темы для обсуждения

- Оценка ассоциаций «генотип-фенотип» и их значимости
- Факторы, влияющие на значимость оценок
- Объединение выборок и метаисследования
- Учет множественности сравнений

Проверка однородности материала и вычисление OR для нескольких выборок

- **Индекс рассеяния** для биномиальных выборок

Можно ли объединить *k* независимых выборок и оценить частоту как

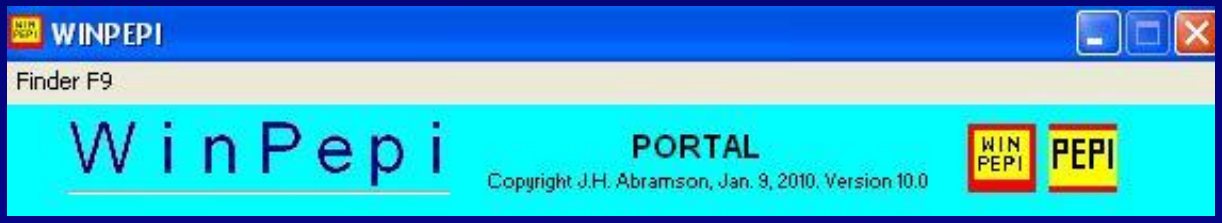
$$\bar{p} = \frac{\sum_i n_i}{\sum_i N_i}$$

Объем выборки	Число мутаций	Частота
N_1	n_1	p_1
N_2	n_2	p_2
….	….	….
N_k	n_k	p_k

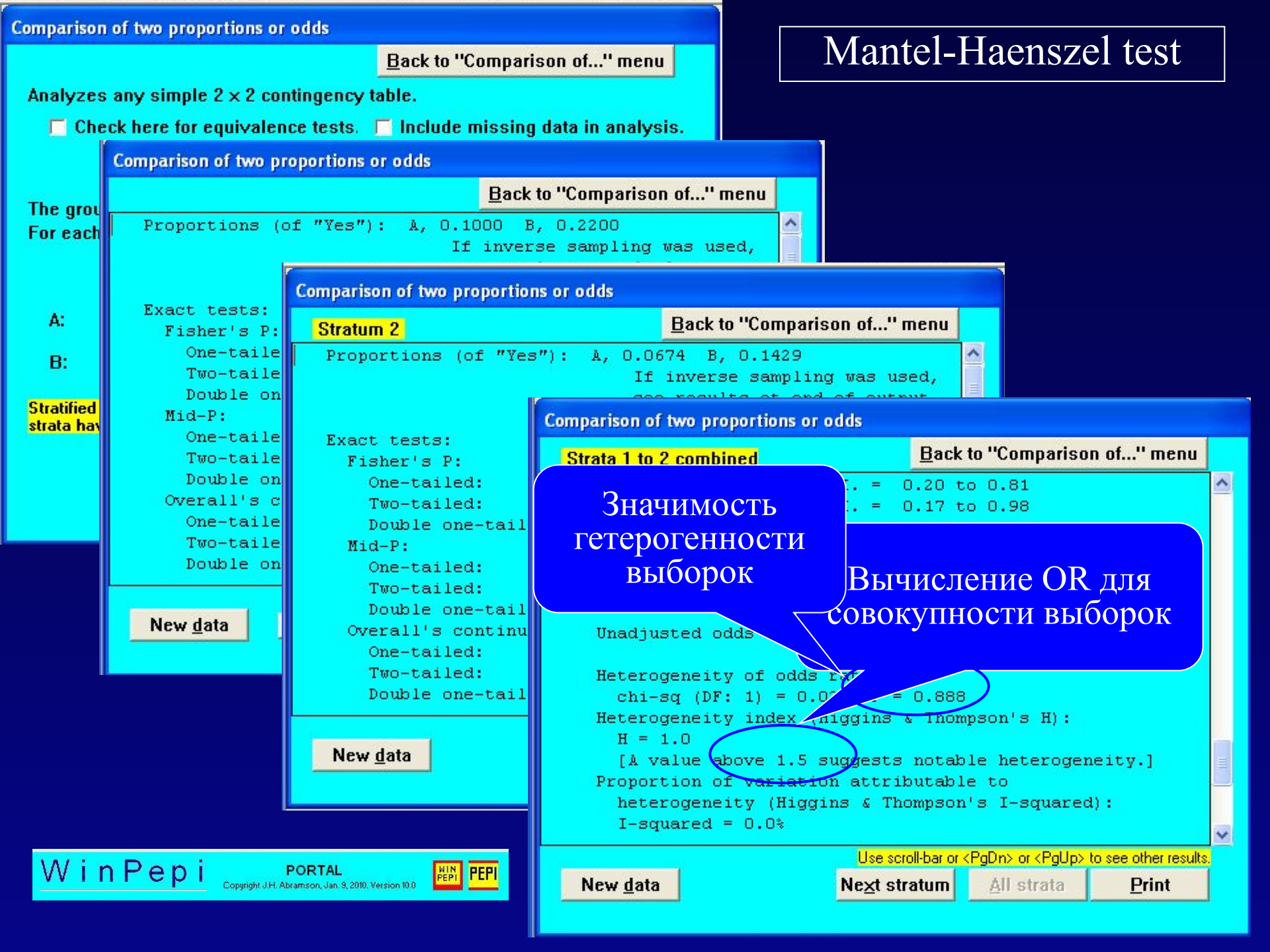
Выборки можно объединять, если

$$\frac{\sum_i N_i (p_i - \bar{p})^2}{\bar{p}} < 2k$$

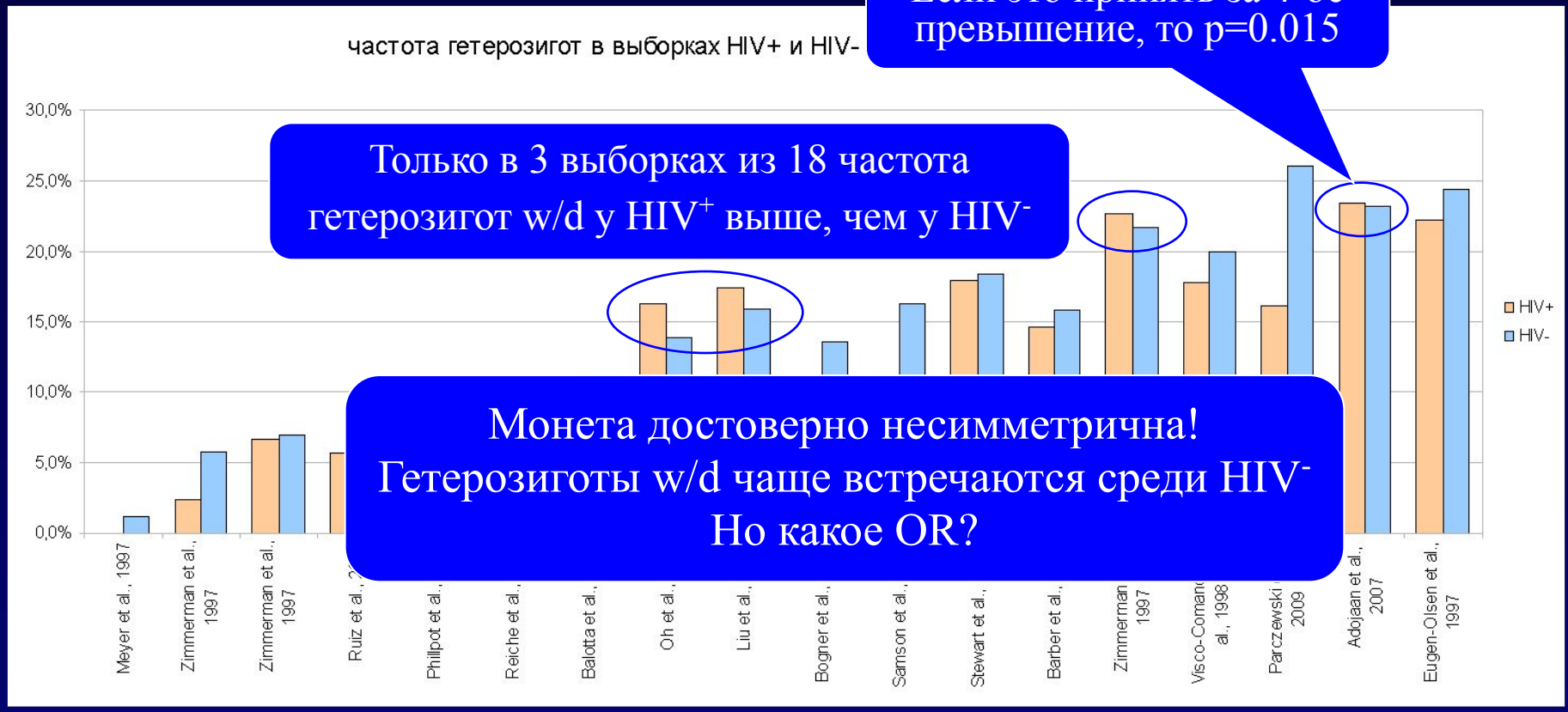
- Mantel-Haenszel test



Mantel-Haenszel test



Объединение выборок с нез[illegible]ами



Если ассоциации нет, то случаи «больше-меньше» должны появляться с вероятностью ½

Вероятность выпадения 3 (и менее) орлов в 18 бросаниях монеты равна

$$p = C_{18}^{3}\left(\frac{1}{2}\right)^{18} + C_{18}^{2}\left(\frac{1}{2}\right)^{18} + C_{18}^{1}\left(\frac{1}{2}\right)^{18} + C_{18}^{0}\left(\frac{1}{2}\right)^{18} \approx 0.0038$$

Ма ты

Протективное действие гетерозиготы w/d CCR5 достоверно, но не велико: OR=1.15

Мета-анализ	OR	RR = f_+ / f_-	Δf = $f_- - f_+$
Mantel-Haenszel оценка	0.87 (1.15)	0.887	0.016
Unadjusted оценка (по всем данным)	0.78	0.801	0.027
95%-довер. интервал	0.77 - 0.97	0.81 - 0.98	0.007-0.023
Значимость гетерогенности (*p*)	0.131	0.236	0.451
Число «null»-статей (OR=1) для ликвидации значимости	7	2	-
Значимость корреляции объемов выборок и эффектов (д.б. > 0.1)	0.188 (Regression asymmetry test, Egger) 0.211 (Adjusted rank correlation, Begg&Mazumdar):		
Итоговая значимость различий (Fisher's two-tailed)	0.014		

Темы для обсуждения

- Оценка ассоциаций «генотип-фенотип» и их значимости
- Факторы, влияющие на значимость оценок
- Объединение выборок и метаисследования
- Учет множественности сравнений

Как это бывает?

Генерируем две одинаково распределенные выборки по 100 особей с 20 некоторыми генотипами
Наблюдаем появление ложных ассоциаций

Частоты минорных аллелей (в среднем 0.1)

Ген	Больные	Здоровые	OR
1	7	8	
2	10	2	
3	17	5	
4	13	12	
5	12	11	
6	7	10	
7	10	12	
8	14	9	
9	14	8	
10	9	12	
11	13	9	
12	9	14	
13	10	13	
14	8	8	
15	14	12	
16	17	7	
17	11	13	
18	10	8	
19	16	10	
20	12	8	

Должно быть OR=1

4

Значим

Сразу 3 локуса «ассоциированы» с заболеваемостью!

Как избежать фальшивых открытий?

- **Правило Карло Бонферрони (1935):**
  При проведение ***m*** независимых статистических тестов значимы только те результаты, для которых

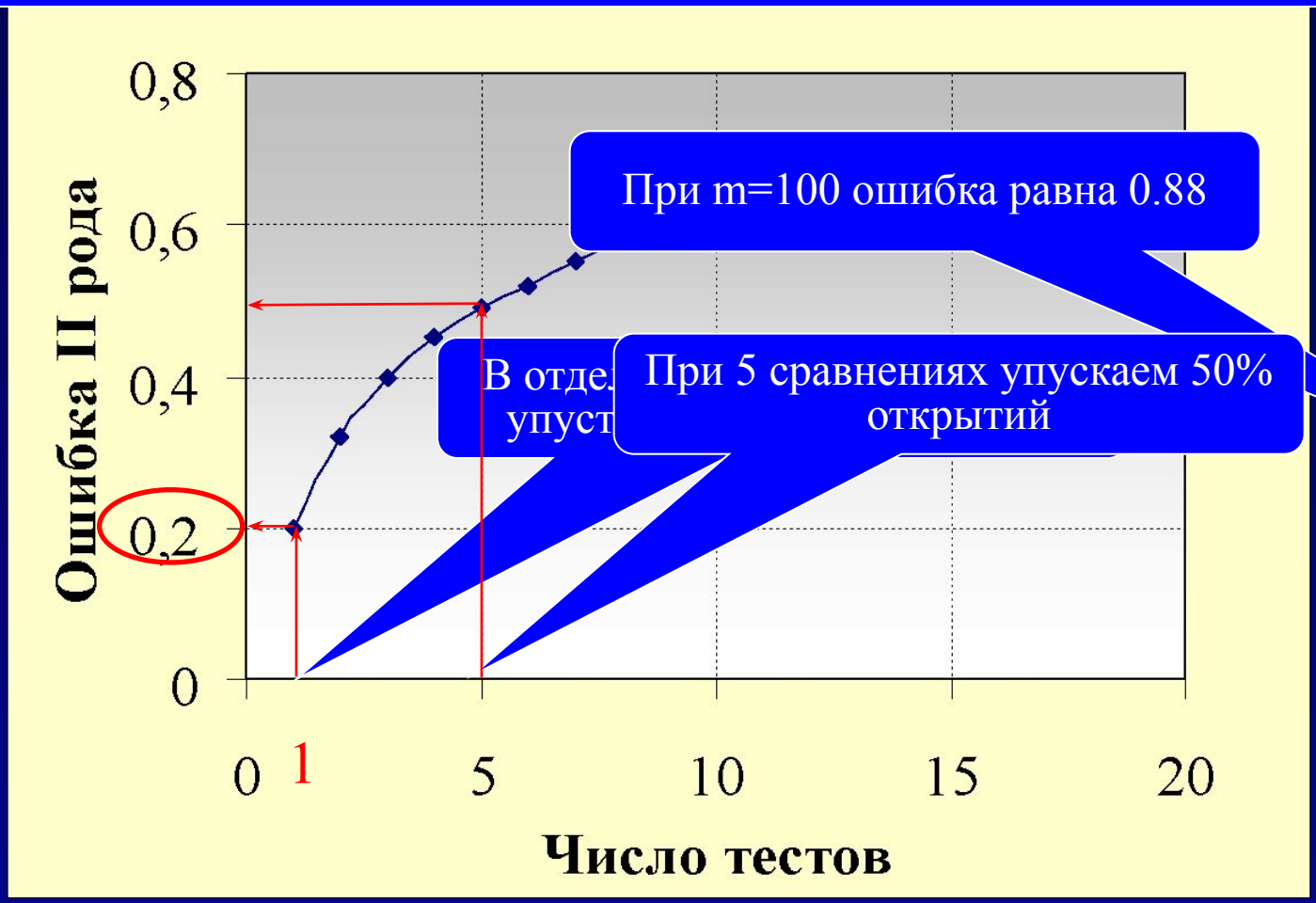
$$p < \frac{0.05}{m}$$

- **False Discovery Rate control**: **FDR** - контроль

- **Permutation test**
  (компьютерная перестановка лэйблов «case-control»)

Зависимость ошибки II рода от числа тестов (SNP)

При 100 сравнениях ради того, чтобы гарантировать отсутствие хотя бы одного ложного результата, мы упускаем 88% открытий!



Новый принцип проверки статистических гипотез: FDR-контроль

False Discovery Rate control: Benjamini, Hochberg (1995)

Вероятность хотя бы одного
фальшивого открытия < Уровня значимости
Ошибка I рода < 0.05

Традиционный принцип заменяется на

105 статей в PubMed Central

Средняя доля фальшивых открытий < Выбранный уровень

$$E\left(\frac{\text{Число неправильно отвергнутых нулевых гипотез}}{\text{Число отвергнутых нулевых гипотез}}\right) < 0.05$$

Пример: множественные сравнения по 10 тестам

Что делать, если FDR не помогает?

Permutation tests:

случайные перестановки пометок «case-control»
в компьютерных симуляциях по алгоритму:

- В исходной базе данных делаем случайную перестановку лейблов case-control

Точный тест Фишера – это тоже permutation test, только реализованный аналитически (p вычисляется
по формулам комбинаторной теории вероятностей)

- Вычисляем откорректированное p как

$$p' = \frac{\text{Число случаев}\left(p_{perm} \le p\right)}{N}$$

Permutation test применительно к данным об ассоциации заболеваемости с 10 SNP

Переставляем отметки «case-control» 10000 раз. В результате получаем коррекцию *p*

SNP	Частота минорного аллеля		OR		
	Case (100)	Control (100)			
1	62	26	4,6	0,0001	0,000
2				0,009	0,010
3			2,8	0,011	0,007
				0,023	0,025
				0,071	0,109
				0,096	0,098
				0,103	0,058
				0,120	0,067
				0,571	0,476
				0,911	1,000

Но так бывает не всегда

Значимо по Бонферрони

Совсем маленькая программка