

# Введение в генетическую эпидемиологию

Ю. С. Аульченко

yurii [dot] aulchenko [at] gmail [dot] com

17 апреля 2012 г.

## Обзор лекции

- 1 Основные эпидемиологические дизайны
- 2 Эпидемиологические исследования бинарных признаков
  - Представление данных в виде таблицы
  - Относительный риск
  - Исследования с случайным выбором
  - Исследования типа "случай-контроль"
- 3 Статистическая значимость
- 4 Заключение

## Оглавление

- 1 Основные эпидемиологические дизайны
- 2 Эпидемиологические исследования бинарных признаков
  - Представление данных в виде таблицы
  - Относительный риск
  - Исследования с случайным выбором
  - Исследования типа "случай-контроль"
- 3 Статистическая значимость
- 4 Заключение

## Эпидемиологические исследования

- Исследования популяций человека, ставящие целью определить детерминанты здоровья и болезней

## Эпидемиологические исследования

- Исследования популяций человека, ставящие целью определить детерминанты здоровья и болезней
- Основным методом исследования является набор и фенотипирование выборок из популяции с последующим статистическим анализом данных с целью выявить факторы риска

## Эпидемиологические исследования

- Исследования популяций человека, ставящие целью определить детерминанты здоровья и болезней
- Основным методом исследования является набор и фенотипирование выборок из популяции с последующим статистическим анализом данных с целью выявить факторы риска
- Генетико-эпидемиологические исследования: исследование вариации генома как фактора риска

## Классификация эпидемиологических исследований

По временному интервалу исследования

- одномоментные (cross-sectional)
- многомоментные (longitudinal)

По способу выбора группы обследуемых

- случайный по отношению к исследуемому признаку (random ascertainment, population-based)
- на основе исследуемого признака (например, case-control)

По наличию других объединяющих характеристик (cohort), например год рождения

## Распространенные типы эпидемиологических исследований

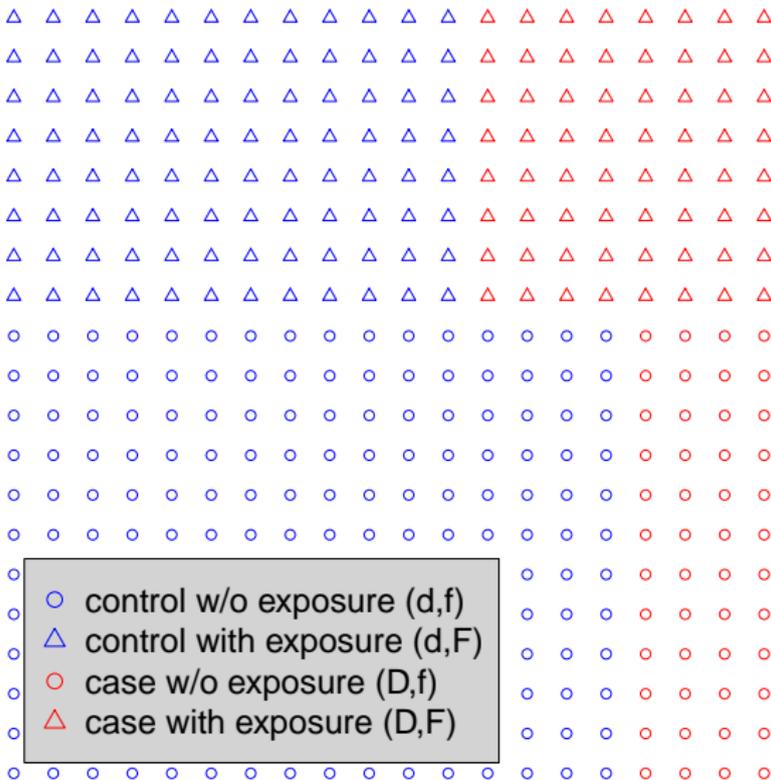
Название	Классификация	Пример
Случай-контроль (case-control)	Одномоментное с выбором по исследуемому признаку	Выборка N больных диабетом из Новосибирска + равное количество "контролей" без диабета
Популяционное когортное (population-based cohort)	Многомоментное проспективное когортное	Rotterdam Study: случайная выборка 11,000 человек возраста 55+ из пригорода Роттердама; первоначальное обследование в 1990-1993, повторное обследование каждые 2-3 года
Семейная одномоментная когорта (cross-sectional family-based cohort)	Одномоментное популяционное семейное	ERF: обследование 3,000 человек из генетически изолированной популяции; все участники являются потомками 21 семьи, живших в XIX веке и имевших $\geq 6$ потомков

## Оглавление

- 1 Основные эпидемиологические дизайны
- 2 Эпидемиологические исследования бинарных признаков
  - Представление данных в виде таблицы
  - Относительный риск
  - Исследования с случайным выбором
  - Исследования типа "случай-контроль"
- 3 Статистическая значимость
- 4 Заключение

Представление данных в виде таблицы

## Генеральная совокупность









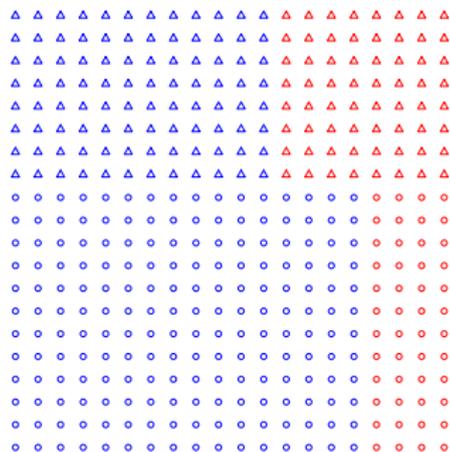






Относительный риск

## Относительный риск

Относительный риск в  
выборке

$$\begin{aligned}
 RR &= \frac{\pi_{D|F}}{\pi_{D|f}} = \frac{64/160}{48/240} = \\
 &= \frac{0.4}{0.2} = 2
 \end{aligned}$$

	d	D	
f	192	48	240
F	96	64	160
	288	112	400



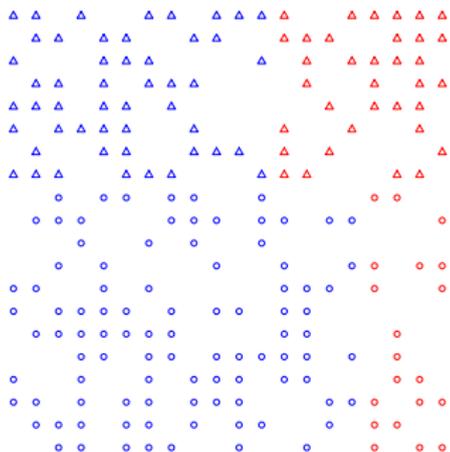
## Характеристики RR как меры ассоциации

- Если фактор не связан с болезнью, ожидается, что  $RR = 1$
- Если фактор более представлен в субпопуляции случаев (повышает риск заболевания),  $RR > 1$
- Если фактор более представлен в субпопуляции контролей (является "протективным"),  $RR < 1$

## Случайная выборка

- Экспериментатор контролирует общий объем выборки ( $N$ )
- Выборка набирается случайным образом из генеральной совокупности
- Вероятность наблюдения определенного класса пропорционально встречаемости этого класса в генеральной совокупности

## Случайная выборка (200 человек)



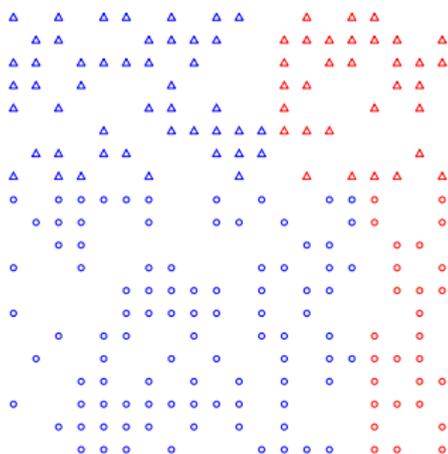
	d	D	
f	95	20	115
F	50	35	85
	145	55	200

Выборочная оценка RR:

$$\begin{aligned}
 RR &= \frac{\pi_{D|F}}{\pi_{D|f}} = \frac{35/85}{20/115} = \\
 &= \frac{0.412}{0.174} = 2.368
 \end{aligned}$$

(оценка выше реальной)

## Ещё одна случайная выборка



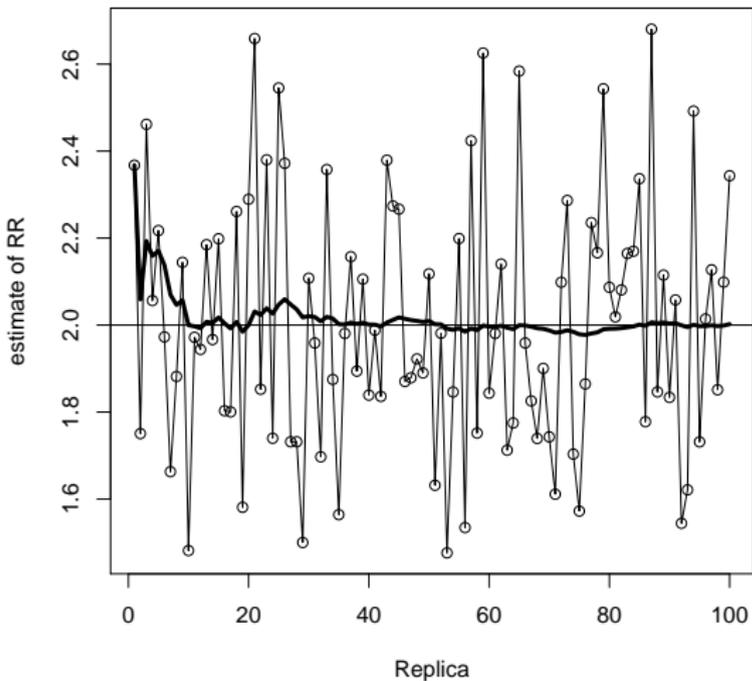
	d	D	
f	93	28	121
F	47	32	79
	140	60	200

Выборочная оценка RR:

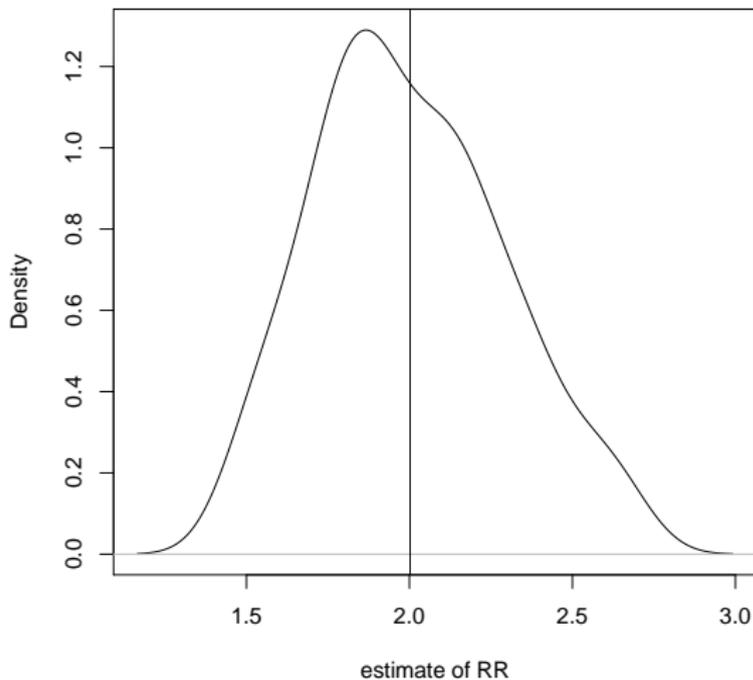
$$\begin{aligned}
 RR &= \frac{\pi_{D|F}}{\pi_{D|f}} = \frac{32/79}{28/121} = \\
 &= \frac{0.405}{0.231} = 1.75
 \end{aligned}$$

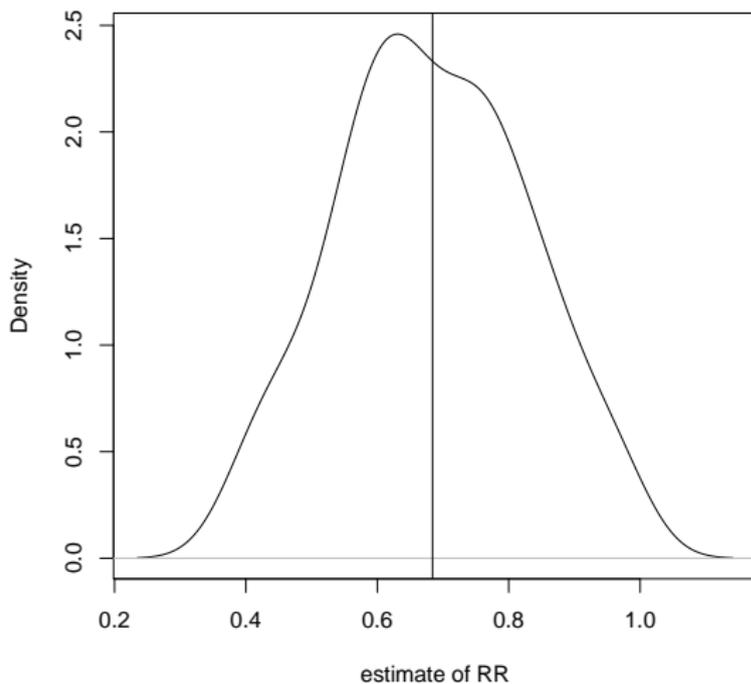
(оценка ниже реальной)

## Однако асимптотически оценка является несмещенной



## Выборочное распределение оценки RR

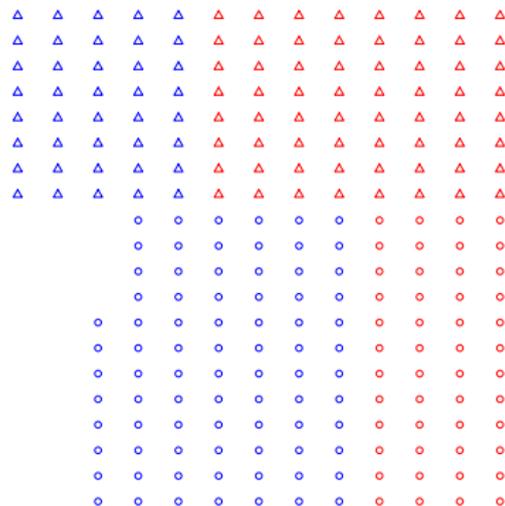


Выборочное распределение оценки  $\log(RR)$ 

## Выборка типа "случай-контроль"

- Экспериментатор контролирует как общий объем выборки ( $N$ ), так и число "случаев" ( $N_D$ ) и "контролей" ( $N_d$ )
- Выборка "случаев" набирается случайным образом из совокупности "случаев"
- Выборка "контролей" набирается случайным образом из совокупности "контролей"
- Такой дизайн позволяет изучать редкие болезни

## Однако... Как измерить силу ассоциации?



	d	D	
f	80	48	128
F	40	64	104
	120	112	232

Эта выборка (112:120)  
сохраняет отношения внутри  
колонок.

$$\begin{aligned}
 (?)RR &= \frac{\pi_{D|F}}{\pi_{D|f}} = \frac{64/104}{48/128} = \\
 &= \frac{0.615}{0.375} = 1.641
 \end{aligned}$$

$$\begin{aligned}
 (!)RR &= \frac{\pi_{D|F}}{\pi_{D|f}} = \\
 &= \frac{64/(40 + 64)}{48/(80 + 48)}
 \end{aligned}$$

## Представление выборки случай-контроль

Выборку можно представить с помощью следующей таблицы

	d	D
f	$N_{f,d} = N_d \cdot (1 - \pi_{F d})$	$N_{f,D} = N_D \cdot (1 - \pi_{F D})$
F	$N_{F,d} = N_d \cdot \pi_{F d}$	$N_{F,D} = N_D \cdot \pi_{F D}$
	$N_d$	$N_D$

- Параметры  $N_d$  и  $N_D$  находятся под контролем исследователя
- Параметры  $\pi_{F|D}$  и  $\pi_{F|d}$  (встречаемость фактора риска в суб-популяциях случаев и контролей) являются характеристикой генеральной совокупности

## Как измерить ассоциацию в выборке случай-контроль?

Необходима мера ассоциации

- Близкая по характеристикам к относительному риску
- Не зависящая от параметров, контролируемых исследователем (конкретного числа случаев и контролей)

## Такой мерой является отношение шансов (OR)

Шансы найти фактор риска

- В субпопуляции случаев:  $N_{F,D}/N_{f,D}$
- В контрольной субпопуляции:  $N_{F,d}/N_{f,d}$

Отношение шансов (odds ratio, OR) в случаях и контролях:

$$\begin{aligned} OR &= \frac{N_{F,D}/N_{f,D}}{N_{F,d}/N_{f,d}} = \frac{N_{F,D} \cdot N_{f,d}}{N_{f,D} \cdot N_{F,d}} = \\ &= \frac{N_D \cdot \pi_{F|D} \cdot N_d \cdot (1 - \pi_{F|d})}{N_D \cdot (1 - \pi_{F|D}) \cdot N_d \cdot \pi_{F|d}} = \frac{\pi_{F|D} \cdot (1 - \pi_{F|d})}{(1 - \pi_{F|D}) \cdot \pi_{F|d}} \end{aligned}$$

не зависит от конкретных  $N_D$  и  $N_d$

## Характеристики OR как меры ассоциации

- Если ассоциации нет,  $OR = 1$
- $OR > 1$  если фактор риска более представлен в субпопуляции случаев
- $OR < 1$  если фактор риска более представлен в субпопуляции контролей

## Характеристики OR как меры ассоциации

Можно показать, что

$$OR = RR \cdot \frac{1 - \pi_{D|f}}{1 - \pi_{D|F}}$$

Таким образом, OR аппроксимирует RR если ...

## Характеристики OR как меры ассоциации

Можно показать, что

$$OR = RR \cdot \frac{1 - \pi_{D|f}}{1 - \pi_{D|F}}$$

Таким образом, OR аппроксимирует RR если ...

- Риск болезни мал как в присутствии, так и в отсутствии фактора риска ( $\pi_{D|f} \rightarrow 0$  и  $\pi_{D|F} \rightarrow 0$ )

## Характеристики OR как меры ассоциации

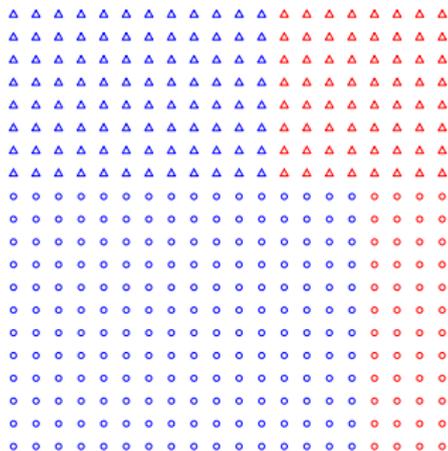
Можно показать, что

$$OR = RR \cdot \frac{1 - \pi_{D|f}}{1 - \pi_{D|F}}$$

Таким образом, OR аппроксимирует RR если ...

- Риск болезни мал как в присутствии, так и в отсутствии фактора риска ( $\pi_{D|f} \rightarrow 0$  и  $\pi_{D|F} \rightarrow 0$ )
- Если риск болезни практически не зависит от исследуемого фактора ( $\pi_{D|f} \approx \pi_{D|F}$ ,  $RR \rightarrow 1$ )

## Отношение шансов в генеральной совокупности



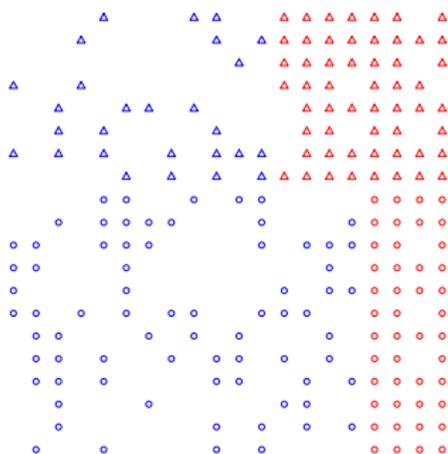
Отношение шансов в  
генеральной совокупности

$$\begin{aligned}
 OR &= \frac{N_{F,D} \cdot N_{f,d}}{N_{f,D} \cdot N_{F,d}} = \\
 &= \frac{64 \cdot 192}{48 \cdot 96} = 2.667
 \end{aligned}$$

	d	D	
f	192	48	240
F	96	64	160
	288	112	400

Случай-контроль

## Выборка "случай-контроль" (100:100)



Выборочная оценка

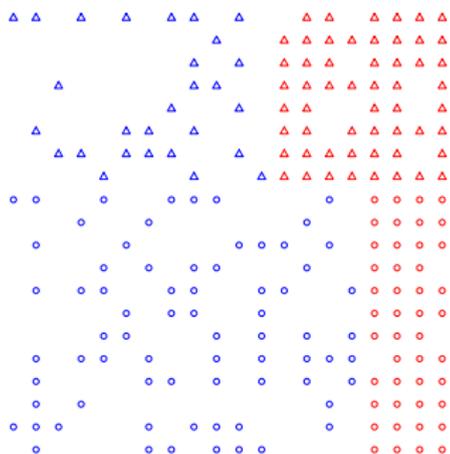
$$\begin{aligned}
 OR &= \frac{N_{F,D} \cdot N_{f,d}}{N_{f,D} \cdot N_{F,d}} = \\
 &= \frac{55 \cdot 73}{45 \cdot 27} = 3.305
 \end{aligned}$$

(оценка выше реальной)

	d	D	
f	73	45	118
F	27	55	82
	100	100	200

Случай-контроль

## Другая выборка 100:100



	d	D	
f	72	45	117
F	28	55	83
	100	100	200

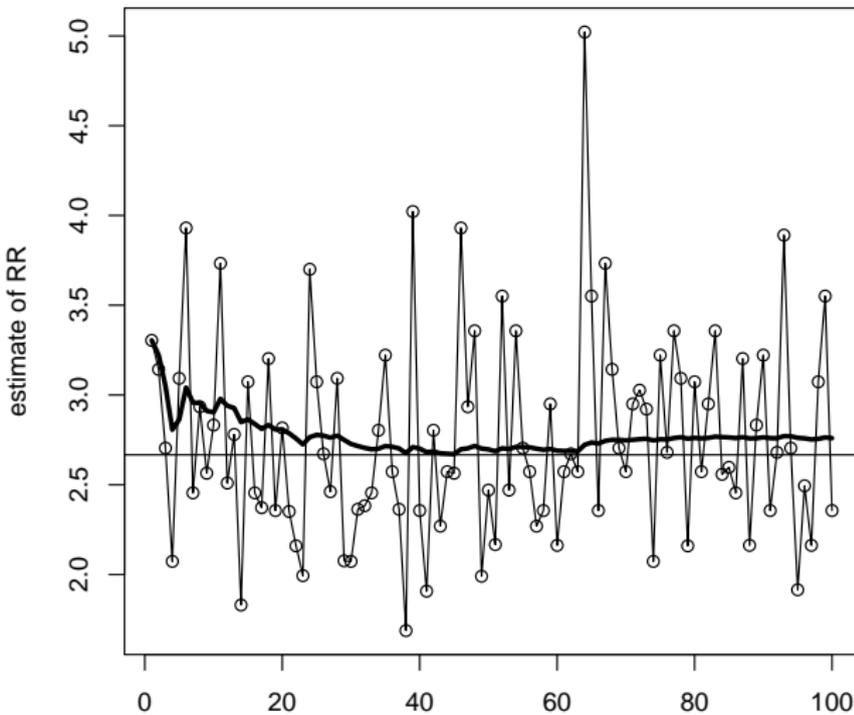
Выборочная оценка

$$\begin{aligned}
 OR &= \frac{N_{F,D} \cdot N_{f,d}}{N_{f,D} \cdot N_{F,d}} = \\
 &= \frac{55 \cdot 72}{45 \cdot 28} = 3.143
 \end{aligned}$$

(оценка выше реальной)

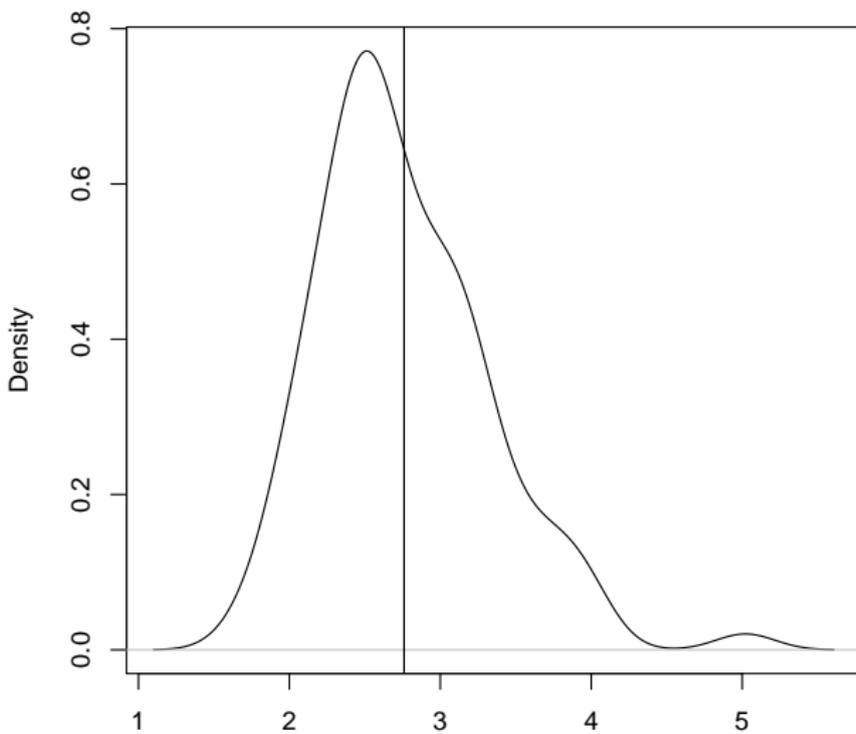
Случай-контроль

## Оценка OR сходится к реальной

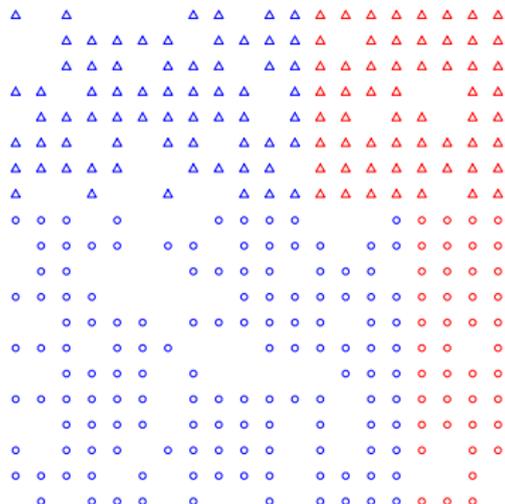


Случай-контроль

## Распределение оценки OR



## Другая выборка (100:200)



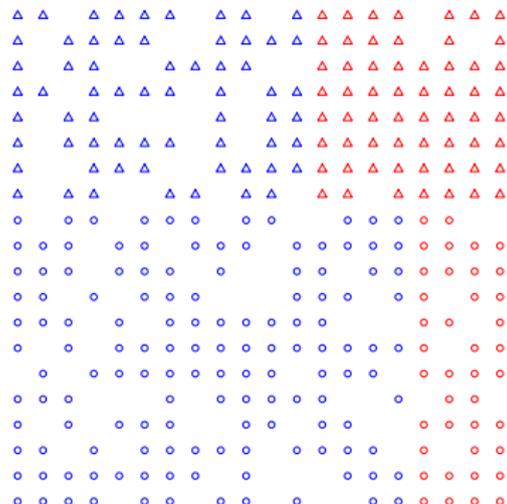
Выборочная оценка

$$\begin{aligned}
 OR &= \frac{N_{F,D} \cdot N_{f,d}}{N_{f,D} \cdot N_{F,d}} = \\
 &= \frac{58 \cdot 133}{42 \cdot 67} = 2.741
 \end{aligned}$$

(оценка выше реальной)

	d	D	
f	133	42	175
F	67	58	125
	200	100	300

## Другая выборка (100:200)



Выборочная оценка

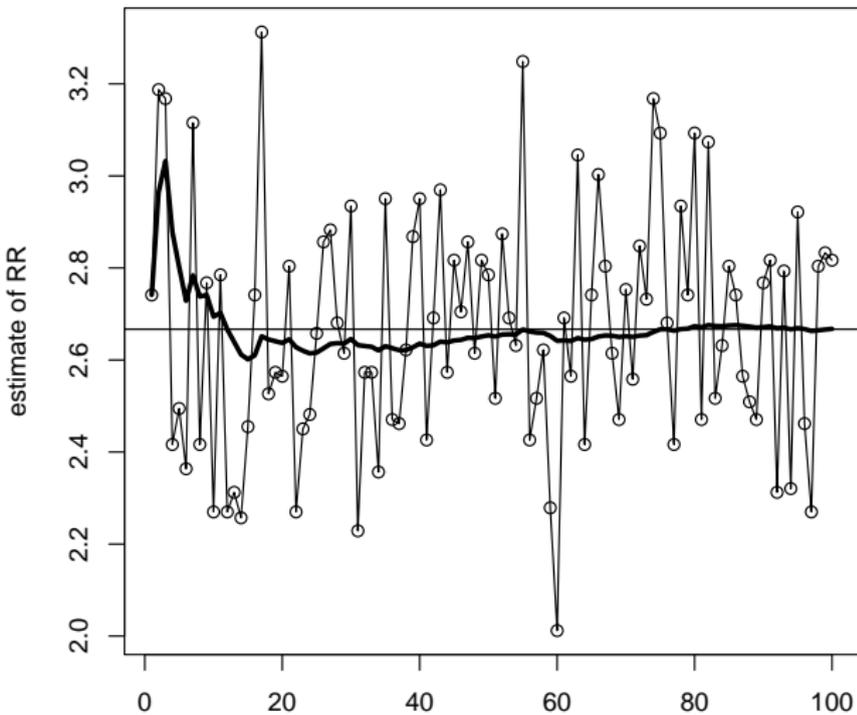
$$\begin{aligned}
 OR &= \frac{N_{F,D} \cdot N_{f,d}}{N_{f,D} \cdot N_{F,d}} = \\
 &= \frac{60 \cdot 136}{40 \cdot 64} = 3.188
 \end{aligned}$$

(оценка выше реальной)

	d	D	
f	136	40	176
F	64	60	124
	200	100	300

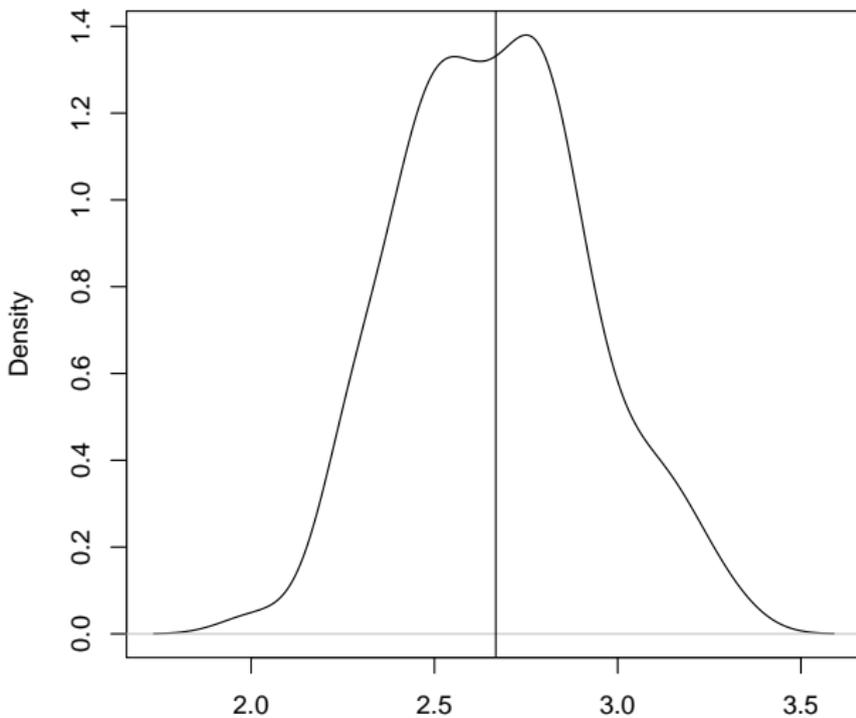
Случай-контроль

## Оценка OR опять сходится



Случай-контроль

## Распределение выборочной оценки OR



## Оглавление

- 1 Основные эпидемиологические дизайны
- 2 Эпидемиологические исследования бинарных признаков
  - Представление данных в виде таблицы
  - Относительный риск
  - Исследования с случайным выбором
  - Исследования типа "случай-контроль"
- 3 Статистическая значимость
- 4 Заключение

## Статистическая значимость ассоциации

Для оценки статистической значимости отклонения RR и OR от единицы можно использовать скор-тест

$$\begin{aligned}T_{x,y}^2 &= n \cdot \hat{\rho}_{xy}^2 = \frac{n \cdot (\text{Cov}(x, y))^2}{\text{Var}(x) \cdot \text{Var}(y)} = \\ &= \frac{n \cdot (\sum(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}\end{aligned}$$

где  $y$  равно 1 для "случаев" и 0 для контролей, а  $x$  равно 1 (наличие фактора риска) или 0 (отсутствие)

## Статистическая значимость ассоциации

Данные для оценки как относительного риска, так и отношения шансов представляются в виде таблицы 2x2

	Control (d)	Case (D)	Total
f	$N_{f,d}$	$N_{f,D}$	$N_f$
F	$N_{F,d}$	$N_{F,D}$	$N_F$
Total	$N_d$	$N_D$	$N$

Можно показать, что в данном случае скор-тест эквивалентен статистике  $\chi^2$  для таблиц сопряженности:

$$T^2 = \sum_{i=\{f,F\}} \sum_{j=\{d,D\}} \frac{(N_{i,j} - E_{i,j})^2}{E_{i,j}}$$

где  $E_{i,j}$  – ожидание численности класса при нулевой гипотезе

## Ожидание численностей классов при нулевой гипотезе

Нулевая гипотеза о независимости вероятности болезни от изучаемого фактора, т. е.

$$E_{i,j} = N \cdot P(i,j) = N \cdot P(i) \cdot P(j),$$

то есть:

- $E_{F,D} = \dots$

## Ожидание численностей классов при нулевой гипотезе

Нулевая гипотеза о независимости вероятности болезни от изучаемого фактора, т. е.

$$E_{i,j} = N \cdot P(i,j) = N \cdot P(i) \cdot P(j),$$

то есть:

- $E_{F,D} = N \cdot P(F) \cdot P(D) = N \cdot (N_F/N) \cdot (N_D/N) = N_F \cdot N_D/N$
- $E_{F,d} = N \cdot P(F) \cdot P(d) = N \cdot (N_F/N) \cdot (N_d/N) = N_F \cdot N_d/N$
- $E_{f,D} = N \cdot P(f) \cdot P(D) = N \cdot (N_f/N) \cdot (N_D/N) = N_f \cdot N_D/N$
- $E_{f,d} = N \cdot P(f) \cdot P(d) = N \cdot (N_f/N) \cdot (N_d/N) = N_f \cdot N_d/N$

## Оглавление

- 1 Основные эпидемиологические дизайны
- 2 Эпидемиологические исследования бинарных признаков
  - Представление данных в виде таблицы
  - Относительный риск
  - Исследования с случайным выбором
  - Исследования типа "случай-контроль"
- 3 Статистическая значимость
- 4 Заключение

## Заключение

- В зависимости от дизайна исследования, ассоциация бинарного признака с бинарным фактором риска измеряется с помощью относительного риска (RR) или отношения шансов (OR) [есть и другие меры]
- Значимость ассоциации можно оценить с помощью скор-теста, который эквивалентен статистике  $\chi^2$  для таблиц сопряженности