

Связь между категориальными переменными

Основы биостатистики, осень 2022

Марина Варфоломеева

- Описание связи между категориальными переменными
- Риск
- Шансы
- Шансы или риск?
- Тест сопряженности хи-квадрат
- Точный критерий Фишера
- G-тест

Описание связи между категориальными переменными

Таблицы сопряженности

Таблицы сопряженности (contingency tables) показывают, как частоты категорий по одной переменной зависят от значения другой категориальной переменной.

Можно узнать:

- Различается ли вероятность быть съеденной птицей у ярких и темных улиток?
- Насколько более вероятно развитие рака легких у курильщиков по сравнению с некурящими?
- Будет ли ниже вероятность инфаркта у людей, принимающих аспирин?

Таблицы сопряженности

Таблицы сопряженности (contingency tables) показывают, как частоты категорий по одной переменной зависят от значения другой категориальной переменной.

Можно узнать:

- Различается ли вероятность быть съеденной птицей у ярких и темных улиток?
- Насколько более вероятно развитие рака легких у курильщиков по сравнению с некурящими?
- Будет ли ниже вероятность инфаркта у людей, принимающих аспирин?

Можно оценить

- относительный риск
- соотношение шансов

Можно протестировать гипотезы о разнице вероятностей

Анализ сопряженности

Анализ сопряженности позволяет оценить, насколько связаны (“сопряжены”) друг с другом категориальные переменные.

Если переменные независимы, то значение одной из переменных не дает информации о вероятностях категорий другой переменной.

Слева — смерти среди 2092 пассажиров Титаника (данные из Dawson, 1995).

Справа — тот же график, если бы вероятность гибели не зависела от пола.

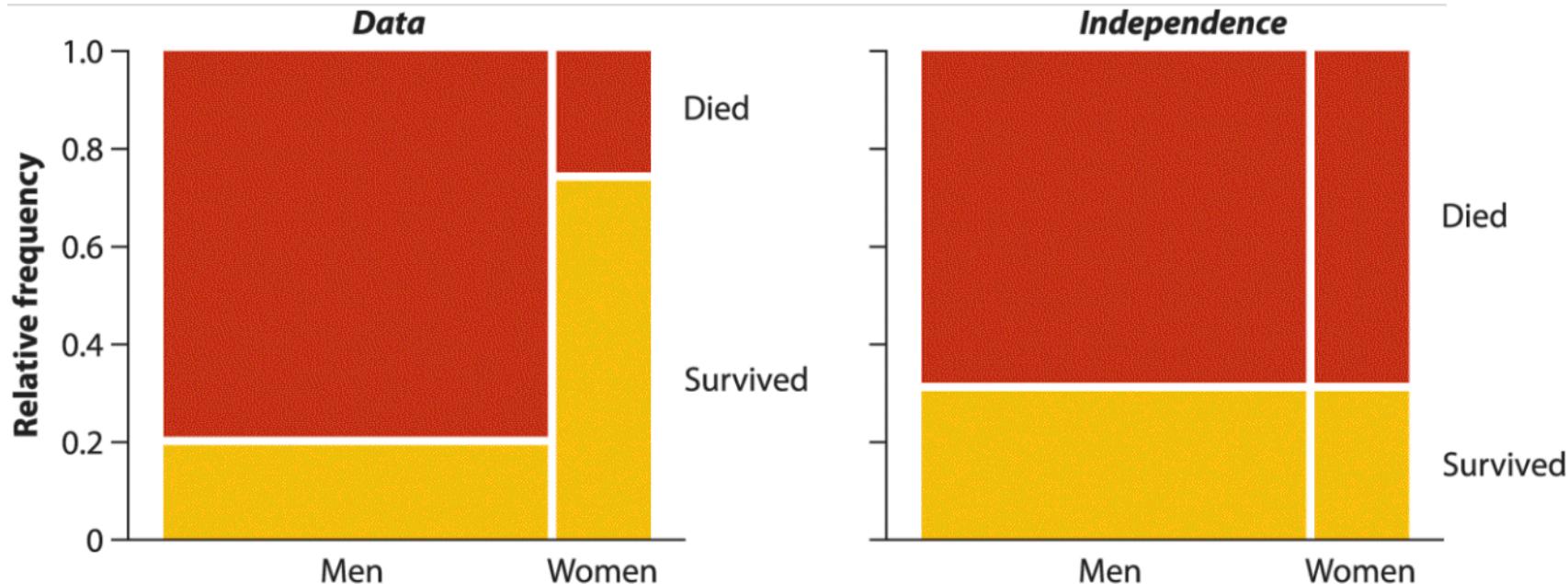


рис. 9.1-1 из Whitlock, Schlüter, 2015

Аспирин и рак

Аспирин используется для лечения головной боли, простуды, профилактики инфаркта и инсульта. В некоторых описательных исследованиях предположили, что он может снижать риск рака.

Экспериментальное слепое исследование (Cook et al. 2005):

- опыт: 19934 женщины — аспирин 100 мг/день
- контроль: 19942 женщины — плацебо

Через 10 лет наблюдений у некоторых
развился рак.

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Аспирин и рак

Аспирин используется для лечения головной боли, простуды, профилактики инфаркта и инсульта. В некоторых описательных исследованиях предположили, что он может снижать риск рака.

Экспериментальное слепое исследование (Cook et al. 2005):

- опыт: 19934 женщины — аспирин 100 мг/день
- контроль: 19942 женщины — плацебо

Через 10 лет наблюдений у некоторых развился рак.

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

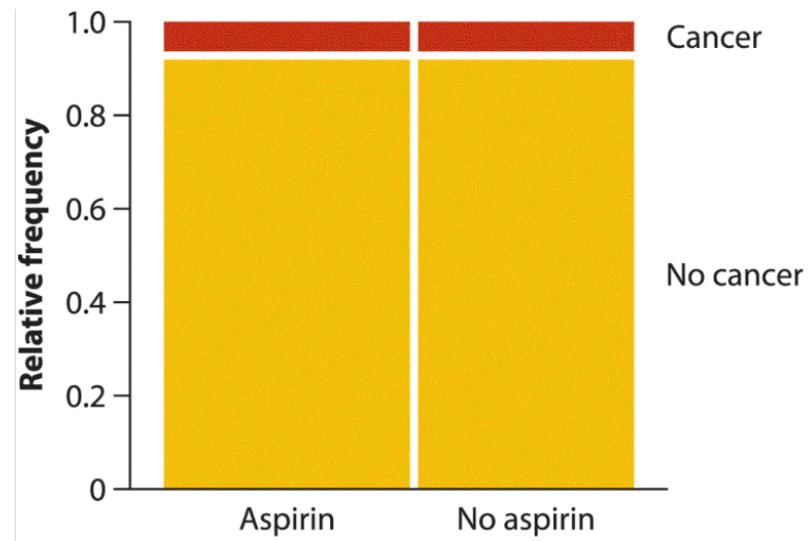


рис 9.2-1 из Whitlock, Schluter, 2015

Аспирин и рак

Аспирин используется для лечения головной боли, простуды, профилактики инфаркта и инсульта. В некоторых описательных исследованиях предположили, что он может снижать риск рака.

Экспериментальное слепое исследование (Cook et al. 2005):

- опыт: 19934 женщины — аспирин 100 мг/день
- контроль: 19942 женщины — плацебо

Через 10 лет наблюдений у некоторых развился рак.

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

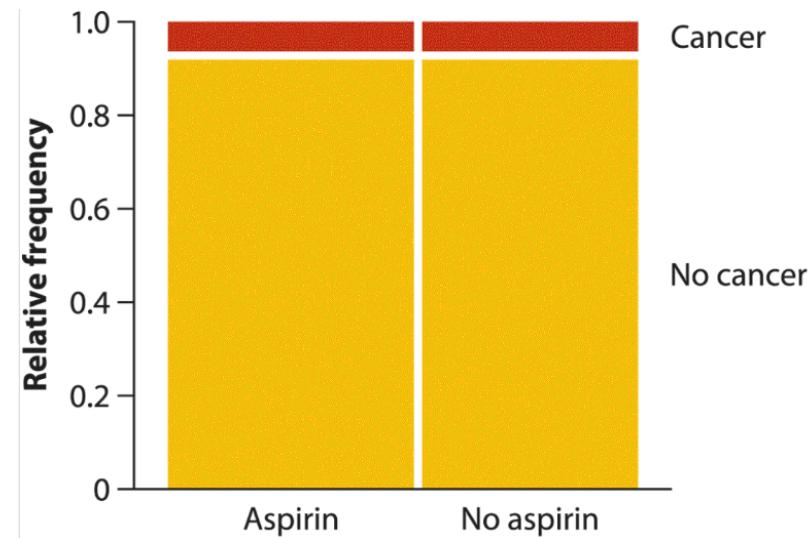


рис 9.2-1 из Whitlock, Schluter, 2015

Можно вычислить вероятности возникновения рака в двух группах. Но хочется описать данные **одним числом**: риск или шансы.

Риск

Риск

Риск (risk) — другое название вероятности определенного исхода.

Риск

Риск (risk) — другое название вероятности определенного исхода.

В примере про аспирин

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Риск

Риск (risk) — другое название вероятности определенного исхода.

В примере про аспирин

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Риск

Аспирин: $p_1 = 1438/19934 = 0.0721$

Плацебо: $p_2 = 1427/19942 = 0.0716$

Относительный риск

Относительный риск (relative risk) — способ сравнения вероятности (риска) определенного исхода (“успеха”) между двумя группами. Отношение выборочных оценок вероятностей этого исхода в сравниваемых группах.

$$RR = \frac{p_1}{p_2}$$

Относительный риск

Относительный риск (relative risk) — способ сравнения вероятности (риска) определенного исхода (“успеха”) между двумя группами. Отношение выборочных оценок вероятностей этого исхода в сравниваемых группах.

$$RR = \frac{p_1}{p_2}$$

$$0 \leq RR < \infty$$

Если $RR = 1$, то риск одинаков в обеих группах.

В примере про аспирин

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

В примере про аспирин

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Риск

Аспирин: $p_1 = 1438/19934 = 0.0721$

Плацебо: $p_2 = 1427/19942 = 0.0716$

В примере про аспирин

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Риск

$$\text{Аспирин: } p_1 = 1438/19934 = 0.0721$$

$$\text{Плацебо: } p_2 = 1427/19942 = 0.0716$$

Относительный риск

$$RR = \frac{0.0721}{0.0716} = 1.007$$

В примере про аспирин

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Риск

$$\text{Аспирин: } p_1 = 1438/19934 = 0.0721$$

$$\text{Плацебо: } p_2 = 1427/19942 = 0.0716$$

Относительный риск

$$RR = \frac{0.0721}{0.0716} = 1.007$$

Т.е. при приеме аспирина относительный риск развития рака даже немножко возрастает по сравнению с плацебо.

В примере про аспирин

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Риск

$$\text{Аспирин: } p_1 = 1438/19934 = 0.0721$$

$$\text{Плацебо: } p_2 = 1427/19942 = 0.0716$$

Относительный риск

$$RR = \frac{0.0721}{0.0716} = 1.007$$

Т.е. при приеме аспирина относительный риск развития рака даже немного возрастает по сравнению с плацебо.

Хотелось бы доверительный интервал.

Стандартная ошибка и доверительный интервал для относительного риска

	опыт	контроль
“успех”	a	b
“неудача”	c	d
Сумма	a + c	b + d

Стандартная ошибка и доверительный интервал для относительного риска

	опыт	контроль
“успех”	a	b
“неудача”	c	d
Сумма	a + c	b + d

Относительный риск $RR = \frac{p_1}{p_2}$ несимметричен
 $0 \leq RR < \infty$

Его логарифм $\ln(RR) = \ln\left(\frac{p_1}{p_2}\right)$ симметричен
 $-\infty \leq \ln(RR) < \infty$

Стандартная ошибка и доверительный интервал для относительного риска

	опыт	контроль
“успех”	a	b
“неудача”	c	d
Сумма	a + c	b + d

Относительный риск $RR = \frac{p_1}{p_2}$ несимметричен
 $0 \leq RR < \infty$

Его логарифм $\ln(RR) = \ln\left(\frac{p_1}{p_2}\right)$ симметричен
 $-\infty \leq \ln(RR) < \infty$

Поэтому сначала делают вычисления в логарифмической шкале:

$$SE_{\ln(RR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{a+c} + \frac{1}{b+d}}$$

$$\ln(RR) - |z| \cdot SE_{\ln(RR)} \leq \ln(RR) \leq \ln(RR) + |z| \cdot SE_{\ln(RR)}$$

Для 95% доверительного интервала $|z_{\text{н.}}| = 1.96$

Стандартная ошибка и доверительный интервал для относительного риска

	опыт	контроль
“успех”	a	b
“неудача”	c	d
Сумма	a + c	b + d

Относительный риск $RR = \frac{p_1}{p_2}$ несимметричен

$$0 \leq RR < \infty$$

Его логарифм $\ln(RR) = \ln\left(\frac{p_1}{p_2}\right)$ симметричен

$$-\infty \leq \ln(RR) < \infty$$

Поэтому сначала делают вычисления в логарифмической шкале:

$$SE_{\ln(RR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{a+c} + \frac{1}{b+d}}$$

$$\ln(RR) - |z| \cdot SE_{\ln(RR)} \leq \ln(RR) \leq \ln(RR) + |z| \cdot SE_{\ln(RR)}$$

Для 95% доверительного интервала $|z_{\text{H.}}| = 1.96$

Потом границы интервала трансформируют обратно в шкалу относительного риска:

$$e^{\ln(RR) - |z| \cdot SE_{\ln(RR)}} \leq RR \leq e^{\ln(RR) + |z| \cdot SE_{\ln(RR)}}$$

Доверительный интервал к риску в примере

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Относительный риск:

$$RR = 1.007$$

Доверительный интервал к риску в примере

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Относительный риск:

$$RR = 1.007$$

Стандартная ошибка:

$$SE_{\ln(RR)} = \sqrt{\frac{1}{1438} + \frac{1}{1427} + \frac{1}{19934} + \frac{1}{19942}} = 0.0387$$

Границы 95% доверительного интервала:

$$\begin{aligned} e^{\ln(1.007) - 1.96 \cdot 0.0387} &\leq RR \leq e^{\ln(1.007) + 1.96 \cdot 0.0387} \\ 0.933 &\leq RR \leq 1.09 \end{aligned}$$

Доверительный интервал к риску в примере

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

Относительный риск:

$$RR = 1.007$$

Стандартная ошибка:

$$SE_{\ln(RR)} = \sqrt{\frac{1}{1438} + \frac{1}{1427} + \frac{1}{19934} + \frac{1}{19942}} = 0.0387$$

Границы 95% доверительного интервала:

$$\begin{aligned} e^{\ln(1.007) - 1.96 \cdot 0.0387} &\leq RR \leq e^{\ln(1.007) + 1.96 \cdot 0.0387} \\ 0.933 &\leq RR \leq 1.09 \end{aligned}$$

Доверительный интервал включает 1. Скорее всего влияние аспирина на риск возникновения рака очень невелико.

Изменение риска

Изменение риска

Снижение относительного риска

(reduction in relative risk) — насколько снижается риск определенного исхода в одной группе по сравнению с другой (с контролем).

$$1 - RR$$

Изменение риска

Снижение относительного риска

(reduction in relative risk) — насколько снижается риск определенного исхода в одной группе по сравнению с другой (с контролем).

$$1 - RR$$

Снижение абсолютного риска (reduction in absolute risk)

— разница абсолютного риска в одной группе (в контроле) по сравнению с другой.

$$p_2 - p_1$$

Изменение риска

Снижение относительного риска

(reduction in relative risk) — насколько снижается риск определенного исхода в одной группе по сравнению с другой (с контролем).

$$1 - RR$$

Иногда бывает, что относительный риск сильно меняется, а абсолютный - ничтожно мало (т.к. очень редкий исход)

Снижение абсолютного риска (reduction in absolute risk)

— разница абсолютного риска в одной группе (в контроле) по сравнению с другой.

$$p_2 - p_1$$

Изменение риска

Снижение относительного риска

(reduction in relative risk) — насколько снижается риск определенного исхода в одной группе по сравнению с другой (с контролем).

$$1 - RR$$

Снижение абсолютного риска (reduction in absolute risk)

— разница абсолютного риска в одной группе (в контроле) по сравнению с другой.

$$p_2 - p_1$$

Иногда бывает, что относительный риск сильно меняется, а абсолютный - ничтожно мало (т.к. очень редкий исход)

При приеме аспирина

Относительное:

$$1 - 1.007 = -0.007$$

Риск развития рака возрастает на 0.007

Изменение риска

Снижение относительного риска

(reduction in relative risk) — насколько снижается риск определенного исхода в одной группе по сравнению с другой (с контролем).

$$1 - RR$$

Снижение абсолютного риска (reduction in absolute risk) — разница абсолютного риска в одной группе (в контроле) по сравнению с другой.

$$p_2 - p_1$$

Иногда бывает, что относительный риск сильно меняется, а абсолютный - ничтожно мало (т.к. очень редкий исход)

При приеме аспирина

Относительное:

$$1 - 1.007 = -0.007$$

Риск развития рака возрастает на 0.007

Абсолютное:

$$0.0716 - 0.0721 = -0.0005$$

Вероятность развития рака возрастает на 0.0005

Шансы

Шансы

Шансы (odds) — это иной способ записи вероятностей: вероятность “успеха”, делённая на вероятность “неудачи”.

Шансы

Шансы (odds) — это иной способ записи вероятностей: вероятность “успеха”, делённая на вероятность “неудачи”.

$$O = \frac{p}{1 - p}$$

$$0 \leq O < \infty$$

Шансы

Шансы (odds) — это иной способ записи вероятностей: вероятность “успеха”, делённая на вероятность “неудачи”.

$$O = \frac{p}{1 - p}$$

$$0 \leq O < \infty$$

Вероятность р	Шансы О	
0.5	1:1 = 1	1 успех : 1 неудача
0.1	1:9 = 0.111	1 успех : 9 неудач
0.0909	1:10 = 0.1	1 успех : 10 неудач

Вероятность

- 0.75
- 0.25
- 0.8
- 0.2

Шансы

Шансы (odds) — это иной способ записи вероятностей: вероятность “успеха”, делённая на вероятность “неудачи”.

$$O = \frac{p}{1 - p}$$

$$0 \leq O < \infty$$

Вероятность р	Шансы О	
0.5	1:1 = 1	1 успех : 1 неудача
0.1	1:9 = 0.111	1 успех : 9 неудач
0.0909	1:10 = 0.1	1 успех : 10 неудач

Вероятность

- 0.75
- 0.25
- 0.8
- 0.2

Шансы

- 3:1
- 1:3
- 4:1
- 1:4

Отношение шансов

Отношение шансов (odds ratio) — используется для сравнения шансов определенного исхода между двумя группами (например, опытом и контролем).

$$OR = \frac{O_1}{O_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

Отношение шансов

Отношение шансов (odds ratio) — используется для сравнения шансов определенного исхода между двумя группами (например, опытом и контролем).

$$OR = \frac{O_1}{O_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

$$0 \leq OR < \infty$$

Если $OR = 1$, то шансы “успеха” одинаковы в обеих группах.

Отношение шансов

Отношение шансов (odds ratio) — используется для сравнения шансов определенного исхода между двумя группами (например, опытом и контролем).

$$OR = \frac{O_1}{O_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

$$0 \leq OR < \infty$$

Если $OR = 1$, то шансы “успеха” одинаковы в обеих группах.

Частоты категорий в двух группах:

	опыт	контроль
“успех”	a	b
“неудача”	c	d

Отношение шансов

Отношение шансов (odds ratio) — используется для сравнения шансов определенного исхода между двумя группами (например, опытом и контролем).

$$OR = \frac{O_1}{O_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

$$0 \leq OR < \infty$$

Если $OR = 1$, то шансы “успеха” одинаковы в обеих группах.

Частоты категорий в двух группах:

	опыт	контроль
“успех”	a	b
“неудача”	c	d

Краткая формула:

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

В примере про аспирин

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515
Сумма	19 934	19 942

частоты категорий

	опыт	контроль
“успех”	a	b
“неудача”	c	d

Риск

Аспирин:

$$p_1 = 1438/19934 = 0.0721$$

Плацебо:

$$p_2 = 1427/19942 = 0.0716$$

Относительный риск:

$$RR = \frac{0.0721}{0.0716} = 1.007$$

Шансы

Аспирин:

$$O_1 = \frac{0.0721}{1 - 0.0721} = 0.0777$$

Плацебо:

$$O_2 = \frac{0.0716}{1 - 0.0716} = 0.0771$$

Отношение шансов:

$$OR = 0.0777/0.0771 = 1.008$$

или $OR = \frac{1438 \cdot 18515}{1427 \cdot 18496} = 1.009$,

т.к. уменьшилась ошибка округления 17 / 47

Стандартная ошибка и доверительный интервал для отношения шансов

	опыт	контроль
“успех”	a	b
“неудача”	c	d
Сумма	a + c	b + d

Стандартная ошибка и доверительный интервал для отношения шансов

	опыт	контроль
“успех”	a	b
“неудача”	c	d
Сумма	a + c	b + d

Отношение шансов $OR = \frac{O_1}{O_2}$ несимметрично
 $0 \leq OR < \infty$

Его логарифм $\ln(OR) = \ln\left(\frac{O_1}{O_2}\right)$ симметричен
 $-\infty \leq \ln(OR) < \infty$

Стандартная ошибка и доверительный интервал для отношения шансов

	опыт	контроль
“успех”	a	b
“неудача”	c	d
Сумма	a + c	b + d

Поэтому сначала делают вычисления в логарифмической шкале:

$$SE_{\ln(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$\ln(OR) - |z| \cdot SE_{\ln(OR)} \leq \ln(OR) \leq \ln(OR) + |z| \cdot SE_{\ln(OR)}$$

Для 95% доверительного интервала $|z_{\text{H.}}| = 1.96$

Отношение шансов $OR = \frac{O_1}{O_2}$ несимметрично
 $0 \leq OR < \infty$

Его логарифм $\ln(OR) = \ln\left(\frac{O_1}{O_2}\right)$ симметричен
 $-\infty \leq \ln(OR) < \infty$

Стандартная ошибка и доверительный интервал для отношения шансов

	опыт	контроль
“успех”	a	b
“неудача”	c	d
Сумма	a + c	b + d

Отношение шансов $OR = \frac{O_1}{O_2}$ несимметрично
 $0 \leq OR < \infty$

Его логарифм $\ln(OR) = \ln\left(\frac{O_1}{O_2}\right)$ симметричен
 $-\infty \leq \ln(OR) < \infty$

Поэтому сначала делают вычисления в логарифмической шкале:

$$SE_{\ln(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$\ln(OR) - |z| \cdot SE_{\ln(OR)} \leq \ln(OR) \leq \ln(OR) + |z| \cdot SE_{\ln(OR)}$$

Для 95% доверительного интервала $|z_{\text{H.}}| = 1.96$

Потом границы интервала трансформируют обратно в шкалу шансов:

$$e^{\ln(OR) - |z| \cdot SE_{\ln(OR)}} \leq OR \leq e^{\ln(OR) + |z| \cdot SE_{\ln(OR)}}$$

Доверительный интервал к отношению шансов в примере

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515

Относительный риск:

$$OR = 1.009$$

Доверительный интервал к отношению шансов в примере

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515

Относительный риск:

$$OR = 1.009$$

Стандартная ошибка:

$$SE_{\ln(RR)} = \sqrt{\frac{1}{1438} + \frac{1}{1427} + \frac{1}{18496} + \frac{1}{18515}} = 0.0388$$

Границы 95% доверительного интервала:

$$\begin{aligned} e^{\ln(1.009) - 1.96 \cdot 0.0388} &\leq RR \leq e^{\ln(1.009) + 1.96 \cdot 0.0388} \\ 0.935 &\leq RR \leq 1.09 \end{aligned}$$

Доверительный интервал к отношению шансов в примере

	Аспирин	Плацебо
Рак	1 438	1 427
Нет рака	18 496	18 515

Относительный риск:

$$OR = 1.009$$

Стандартная ошибка:

$$SE_{\ln(RR)} = \sqrt{\frac{1}{1438} + \frac{1}{1427} + \frac{1}{18496} + \frac{1}{18515}} = 0.0388$$

Границы 95% доверительного интервала:

$$\begin{aligned} e^{\ln(1.009) - 1.96 \cdot 0.0388} &\leq RR \leq e^{\ln(1.009) + 1.96 \cdot 0.0388} \\ 0.935 &\leq RR \leq 1.09 \end{aligned}$$

Доверительный интервал включает 1. Скорее всего влияние аспирина на шансы возникновения рака крайне невелико.

Шансы или риск?

Шансы или риск?

Обе меры используются в биологии

Говорят, что RR более интуитивно-понятен

$OR \approx RR$ когда вероятность “успеха” в целом низка

Выбор зависит от дизайна исследования!

Токсоплазма и автомобильные аварии

Toxoplasma gondii — это паразитический протист, заражающий мозг птиц и млекопитающих и влияющий на их поведение. 25% людей инфицированы токсоплазмой.

Зараженность токсоплазмой в выборках водителей 21-40 лет, попадавших в автомобильные аварии и без истории аварий. (Yereli et al., 2006)

Связан ли токсоплазмоз на вероятность попадания в аварию?

	Инфекция	Нет инфекции
Водители с авариями	61	124
Водители без аварий	16	169

Токсоплазма и автомобильные аварии

Toxoplasma gondii — это паразитический протист, заражающий мозг птиц и млекопитающих и влияющий на их поведение. 25% людей инфицированы токсоплазмой.

Зараженность токсоплазмой в выборках водителей 21-40 лет, попадавших в автомобильные аварии и без истории аварий. (Yereli et al., 2006)

Связан ли токсоплазмоз на вероятность попадания в аварию?

	Инфекция	Нет инфекции
Водители с авариями	61	124
Водители без аварий	16	169

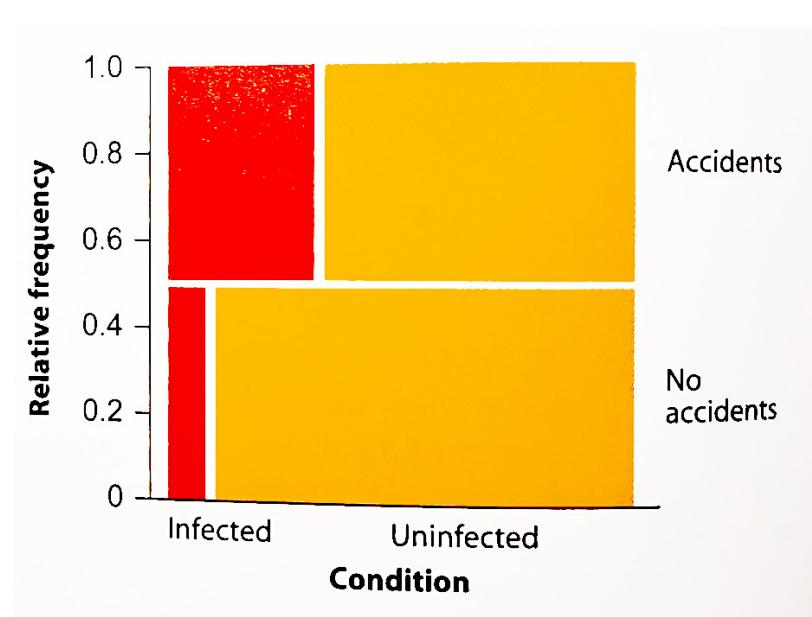


рис.9.3-1 из Whitlock, Schluter, 2015

Пример про токсоплазму

	Инфекция	Нет инфекции
Водители с авариями	61	124
Водители без аварий	16	169

частоты категорий

	группа 1	группа 2
“успех”	a	b
“неудача”	c	d

Риск

Вероятность аварии нельзя сосчитать

Относительный риск не оценить

Шансы

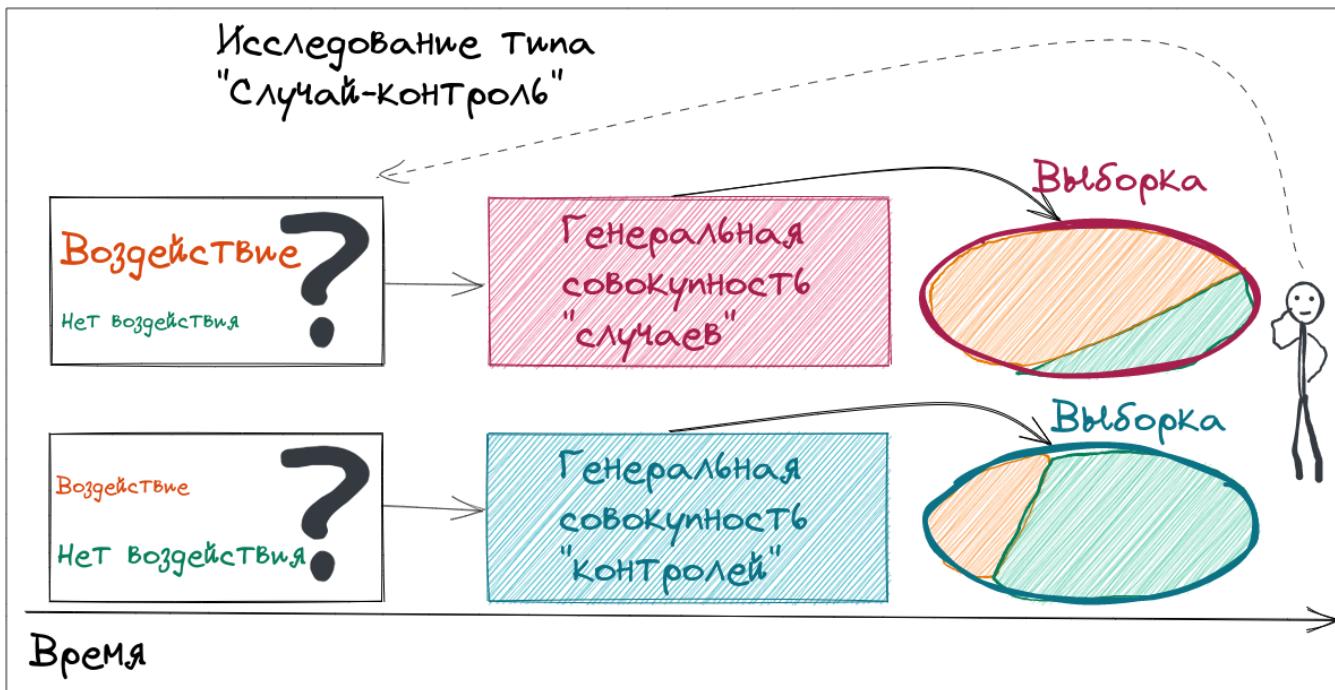
Отношение шансов можно оценить

$$OR = \frac{61 \cdot 169}{16 \cdot 124} = 5.2$$

Исследования типа случай-контроль

(case-control studies)

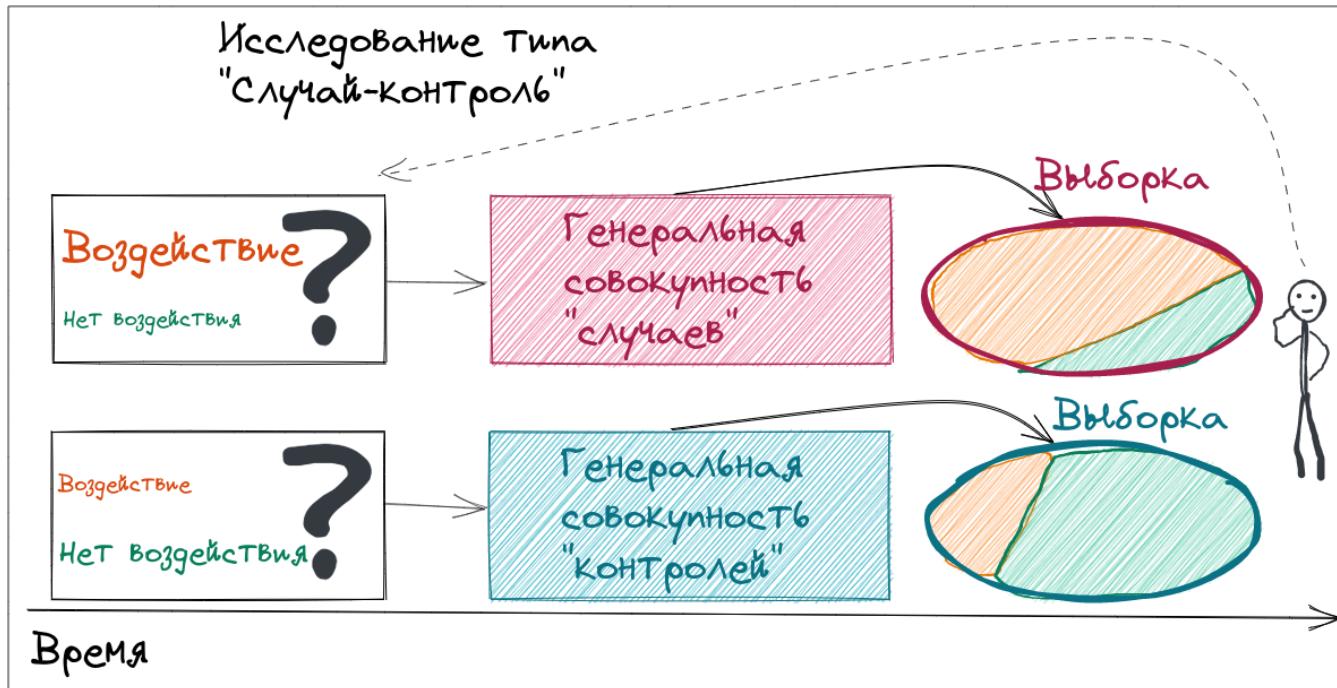
- спланированные описательные исследования
- случайная выборка “случаев” сравнивается с выборкой “контролей”
- оценивают долю субъектов под воздействием среди “случаев” и “контролей”
- хорошо для исследования редких болезней
- нельзя считать RR , т.к. соотношение случай:контроль задано исследователем
- можно считать OR



Исследования типа случай-контроль

(case-control studies)

- спланированные описательные исследования
- случайная выборка “случаев” сравнивается с выборкой “контролей”
- оценивают долю субъектов под воздействием среди “случаев” и “контролей”
- хорошо для исследования редких болезней
- нельзя считать RR , т.к. соотношение случай:контроль задано исследователем
- можно считать OR

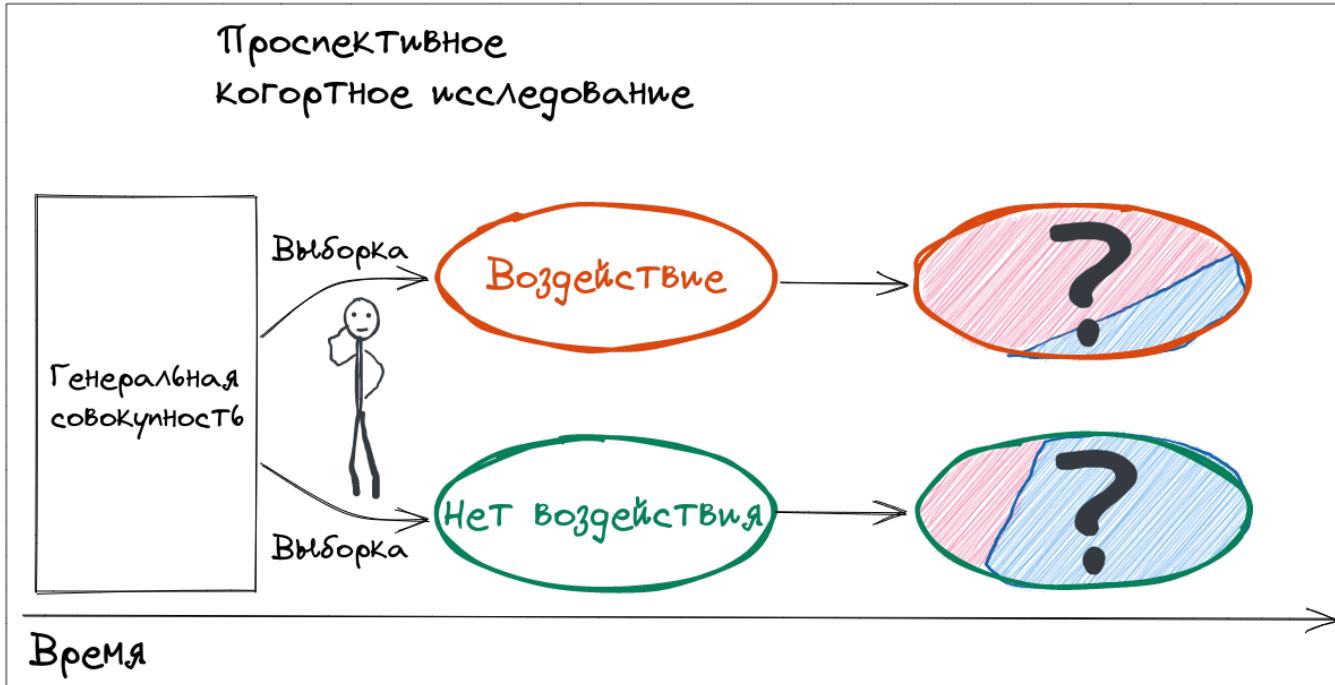


Пример про токсоплазму

Проспективные когортные исследования

(prospective cohort studies)

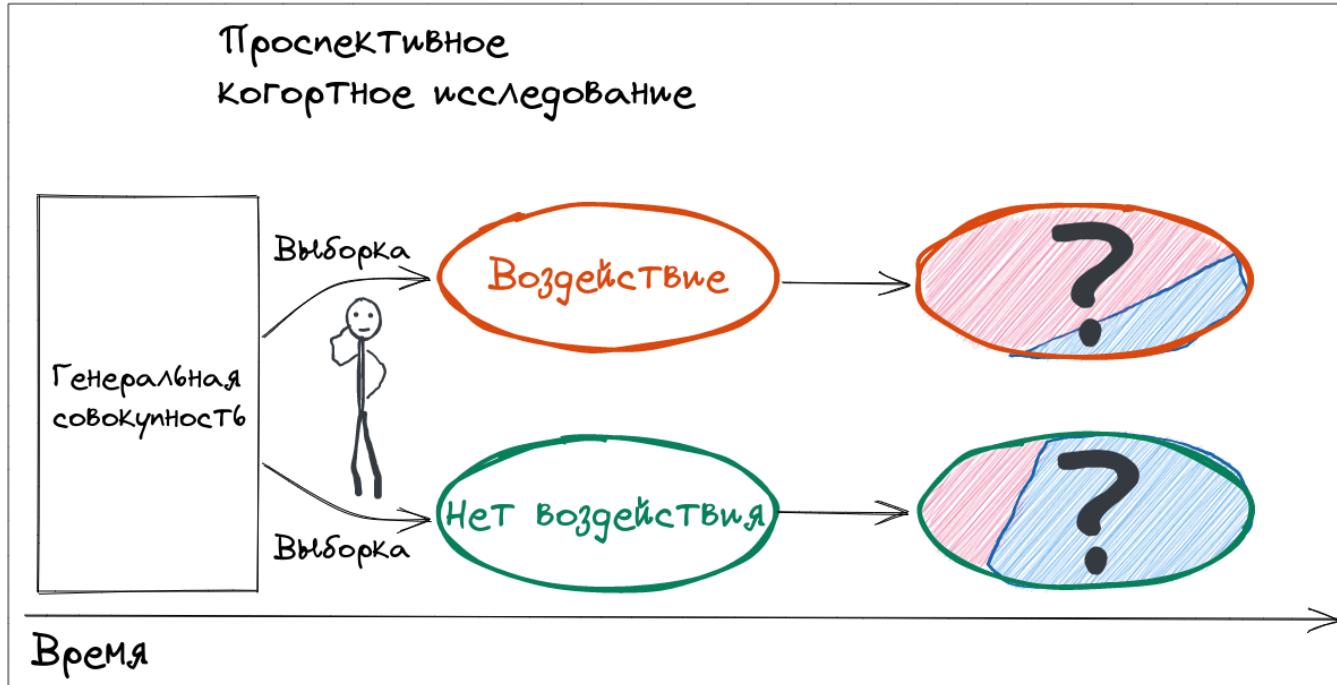
- спланированные описательные исследования
- информация о субъектах (и о воздействии) собрана в начале исследования
- судьбу субъектов прослеживают до наступления “исхода”
- оценивают вероятность наступления “исхода” в зависимости от наличия воздействия
- используются, например, для тестирования новых лекарств и методов лечения
- можно считать и RR , и OR



Проспективные когортные исследования

(prospective cohort studies)

- спланированные описательные исследования
- информация о субъектах (и о воздействии) собрана в начале исследования
- судьбу субъектов прослеживают до наступления “исхода”
- оценивают вероятность наступления “исхода” в зависимости от наличия воздействия
- используются, например, для тестирования новых лекарств и методов лечения
- можно считать и RR , и OR

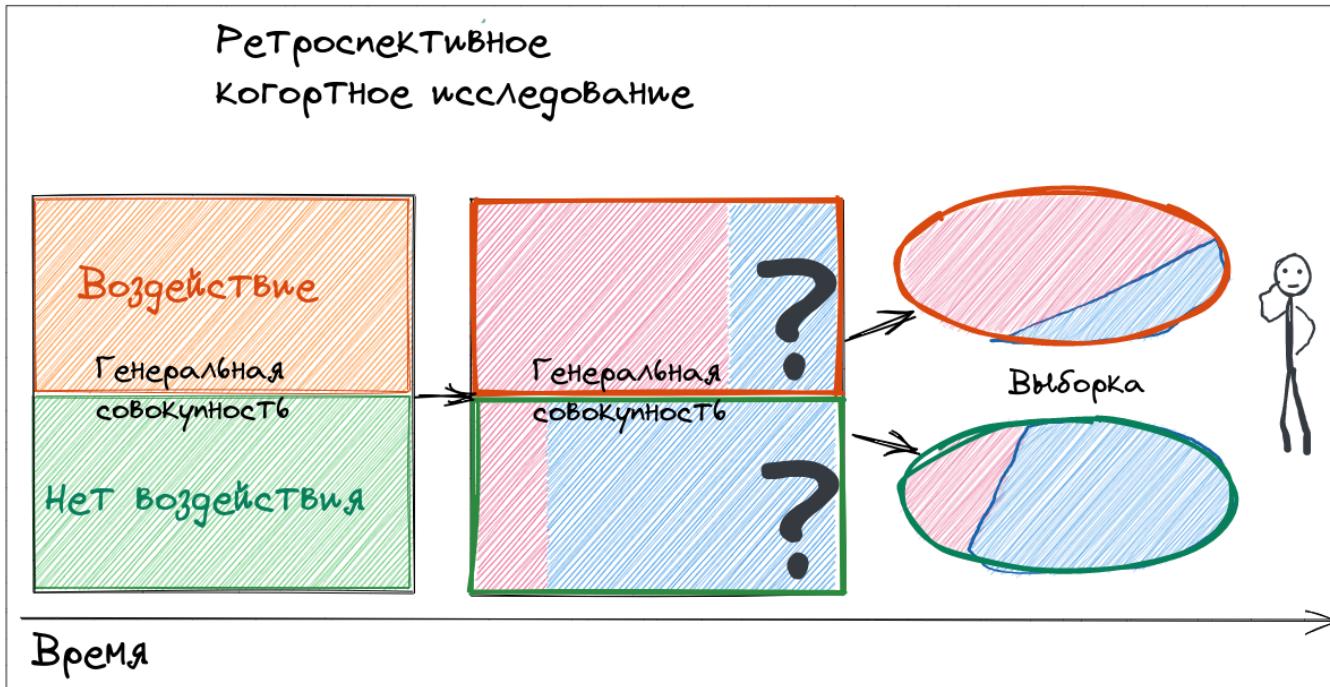


Пример про аспирин

Ретроспективные когортные исследования

(retrospective cohort studies)

- **не** запланированные описательные исследования
- информацию о субъектах (и о воздействии) на них собирают уже после того, как наступил (или нет) "исход"
- оценивают вероятность наступления "исхода" в зависимости от воздействия
- используются, например, для поиска и оценки потенциальных факторов риска
- можно считать и RR , и OR



Тест сопряженности хи-квадрат

Паразиты рыб: “передай другому”

У trematod *Euhaplorchis californiensis* три хозяина в жизненном цикле: улитка, рыба и птица. Инфицированные рыбы проводят много времени у поверхности воды могут стать добычей птиц (Lafferty, Morris, 1996).

Влияет ли уровень заражения trematodами на вероятность поедания птицами?

	Нет заражения	Низкое	Высокое	Сумма
Съедены	1	10	37	48
Не съедены	49	35	9	93
Сумма	50	45	46	141

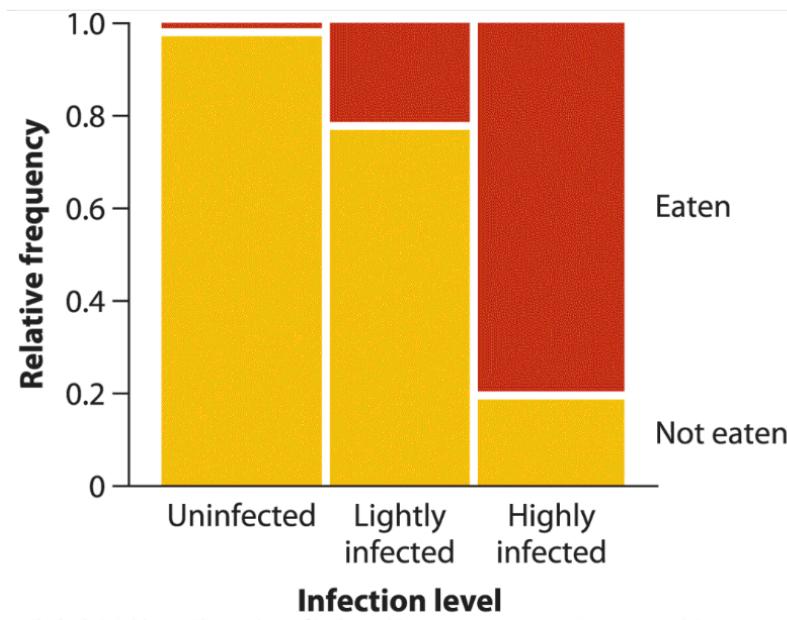


рис. 9.4-1 из Whitlock, Schluter, 2015

Тест сопряженности хи-квадрат

χ^2 -тест позволяет протестировать гипотезу о независимости двух категориальных переменных.

H_0 : — категориальные переменные независимы друг от друга

H_A : — категориальные переменные зависимы

Тест сопряженности хи-квадрат

χ^2 -тест позволяет протестировать гипотезу о независимости двух категориальных переменных.

H_0 : — категориальные переменные независимы друг от друга

H_A : — категориальные переменные зависимы

В примере:

H_0 : — будет ли съедена улитка птицей не зависит от уровня заражения улитки trematodами

H_A : — поедание улитки птицей зависит от уровня заражения улитки trematодами

Ожидаемые частоты в teste сопряженности хи-квадрат

Если переменные независимы (при H_0), вероятность попадания в какую-то категорию по каждой из переменных равна произведению вероятностей этих категорий.

$$P(\text{row}, \text{col}) = P(\text{row}) \cdot P(\text{col}) = \frac{N_{\text{row}}}{N} \cdot \frac{N_{\text{col}}}{N}$$

Чтобы получить частоты, умножаем на общее число наблюдений. Т.е. коротко:

$$\text{Expected}(\text{row}, \text{col}) = \frac{N_{\text{row}} N_{\text{col}}}{N}$$

Ожидаемые частоты в teste сопряженности хи-квадрат

Если переменные независимы (при H_0), вероятность попадания в какую-то категорию по каждой из переменных равна произведению вероятностей этих категорий.

$$P(\text{row}, \text{col}) = P(\text{row}) \cdot P(\text{col}) = \frac{N_{\text{row}}}{N} \cdot \frac{N_{\text{col}}}{N}$$

Чтобы получить частоты, умножаем на общее число наблюдений. Т.е. коротко:

$$\text{Expected}(\text{row}, \text{col}) = \frac{N_{\text{row}} N_{\text{col}}}{N}$$

Наблюдаемые частоты

	Нет заражения	Низкое	Высокое	Сумма
Съедены	1	10	37	48
Не съедены	49	35	9	93
Сумма	50	45	46	141

Ожидаемые частоты

	Нет заражения	Низкое	Высокое	Сумма
Съедены	17.0	15.3	15.7	48
Не съедены	33.0	29.7	30.3	93
Сумма	50	45	46	141

Хи-квадрат статистика для таблиц сопряженности

$$\chi^2 = \sum_{row=1}^r \sum_{col=1}^c \frac{(Observed_{(row,col)} - Expected_{(row,col)})^2}{Expected_{(row,col)}}$$

$$df = (r - 1)(c - 1)$$

Односторонний тест

Наблюдаемые частоты

	Нет заражения	Низкое	Высокое	Сумма
Съедены	1	10	37	48
Не съедены	49	35	9	93
Сумма	50	45	46	141

$$\chi^2 = \frac{(1-17.9)^2}{17.0} + \frac{(49-33.0)^2}{33.0} + \dots = 69.5$$

$$df = (2 - 1)(3 - 1) = 2$$

$$p = 7.77e - 16$$

Ожидаемые частоты

	Нет заражения	Низкое	Высокое	Сумма
Съедены	17.0	15.3	15.7	48
Не съедены	33.0	29.7	30.3	93
Сумма	50	45	46	141

Хи-квадрат статистика для таблиц сопряженности

$$\chi^2 = \sum_{row=1}^r \sum_{col=1}^c \frac{(Observed_{(row,col)} - Expected_{(row,col)})^2}{Expected_{(row,col)}}$$

$$df = (r - 1)(c - 1)$$

Односторонний тест

Наблюдаемые частоты

	Нет заражения	Низкое	Высокое	Сумма
Съедены	1	10	37	48
Не съедены	49	35	9	93
Сумма	50	45	46	141

$$\chi^2 = \frac{(1-17.9)^2}{17.0} + \frac{(49-33.0)^2}{33.0} + \dots = 69.5$$

$$df = (2 - 1)(3 - 1) = 2$$

$$p = 7.77e - 16$$

Вероятность поедания птицей статистически значимо зависит от уровня заражения улитки trematodами.

Ожидаемые частоты

	Нет заражения	Низкое	Высокое	Сумма
Съедены	17.0	15.3	15.7	48
Не съедены	33.0	29.7	30.3	93
Сумма	50	45	46	141

Условия применимости хи-квадрат теста сопряженности

χ^2 -тест для таблиц сопряженности — это частный случай χ^2 -теста адекватности модели, поэтому условия применимости такие же.

- наблюдения независимы друг от друга

χ^2 -статистика приблизительно следует χ^2 -распределению, если:

- нет ожидаемых частот < 1
- $\leq 20\%$ ожидаемых частот < 5

Условия применимости хи-квадрат теста сопряженности

χ^2 -тест для таблиц сопряженности — это частный случай χ^2 -теста адекватности модели, поэтому условия применимости такие же.

- наблюдения независимы друг от друга

χ^2 -статистика приблизительно следует χ^2 -распределению, если:

- нет ожидаемых частот < 1
- $\leq 20\%$ ожидаемых частот < 5

Если условия нарушены:

- Если таблица больше чем 2×2 , можно **объединить редкие категории**, если они имеют биологический смысл
- Если таблица 2×2 , можно использовать **точный тест Фишера**
- Можно использовать **пермутационный тест**

Поправка на непрерывность

Поправка Йейтса на непрерывность (Yates correction for continuity) — используется в анализе таблиц сопряженности 2x2. Корректирует ошибку в результате аппроксимации дискретных вероятностей категорий непрерывным распределением χ^2 .

$$\chi^2 = \sum_{row=1}^r \sum_{col=1}^c \frac{\left(|Observed_{(row,col)} - Expected_{(row,col)}| - \frac{1}{2} \right)^2}{Expected_{(row,col)}}$$

Не рекомендуется.

χ^2 -тест сопряженности с поправкой Йейтса становится слишком консервативным (Maxwell, 1976): значения р завышены.

Точный критерий Фишера

Точный критерий Фишера

Точный критерий Фишера (Fisher's exact test) — тест для таблиц сопряженности 2x2

H_0 : — категориальные переменные независимы друг от друга

H_A : — категориальные переменные зависимы

- даёт точное значение р
- работает и с малыми ожидаемыми частотами

Вручную считать сложно.

Питание вампиров

Летучие мыши-вампиры *Desmodus rotundus* в Коста Рике часто питаются кровью крупного рогатого скота. Кажется, они предпочитают коров быкам, и возможно, реагируют на гормоны.

Влияет ли эструс коров на вероятность быть укушенной вампиром (Turner, 1975)?

	Эструс	Нет эструса	Сумма
Укушена	15	6	21
Не укушена	7	322	329
Сумма	22	328	350



blog.seniorennet.be, CC0, via Wikimedia Commons

Хи-квадрат не подходит для этих данных

Наблюдаемые частоты

	Эструс	Нет эструса	Сумма
Укушена	15	6	21
Не укушена	7	322	329
Сумма	22	328	350

Ожидаемые частоты

	Эструс	Нет эструса	Сумма
Укушена	1.3	19.7	21
Не укушена	20.7	308.3	329
Сумма	22	328	350

Хи-квадрат не подходит для этих данных

Наблюдаемые частоты

	Эструс	Нет эструса	Сумма
Укушена	15	6	21
Не укушена	7	322	329
Сумма	22	328	350

Ожидаемые частоты

	Эструс	Нет эструса	Сумма
Укушена	1.3	19.7	21
Не укушена	20.7	308.3	329
Сумма	22	328	350

Данные не подходят для анализа при помощи хи-квадрат т.к. $\sim 1/4 > 20\%$ ожидаемых частот < 5

Точный критерий Фишера

Наблюдаемые частоты

	Эструс	Нет эструса
Укушена	15	6
Не укушена	7	322

Сколько возможно еще более “экстремальных” (менее вероятных) таблиц?

Суммы по строкам и столбцам не должны при этом меняться.

Точный критерий Фишера

Наблюдаемые частоты

	Эструс	Нет эструса
Укушена	15	6
Не укушена	7	322

Сколько возможно еще более “экстремальных” (менее вероятных) таблиц?

Суммы по строкам и столбцам не должны при этом меняться.

16	5
6	323

17	4
5	324

18	3
4	325

Точный критерий Фишера
учтёт вероятности **всех**
возможных более
экстремальных таблиц.

19	2
3	326

20	1
2	327

21	0
1	328

Whitlock, Schluter, 2015

В нашем примере $p < 10^{-10}$.

Точный критерий Фишера

Наблюдаемые частоты

	Эструс	Нет эструса
Укушена	15	6
Не укушена	7	322

Сколько возможно еще более “экстремальных” (менее вероятных) таблиц?

Суммы по строкам и столбцам не должны при этом меняться.

16	5
6	323

17	4
5	324

18	3
4	325

Точный критерий Фишера
учтёт вероятности **всех**
возможных более
экстремальных таблиц.

19	2
3	326

20	1
2	327

21	0
1	328

Whitlock, Schluter, 2015

В нашем примере $p < 10^{-10}$.

У коров вероятность быть укушенной вампиром статистически-значимо связана с эструсом.

Особенности точного критерия Фишера

- Чрезмерно консервативен (альтернатива - тест Бошлу, Boschloo's test)
- Не так уж и “точен”, т.к. используются вероятности таблиц при зафиксированных значениях сумм по строкам и столбцам (а они могут меняться в выборках).
- Не подходит для стратифицированных данных (тест Коクрана-Мантела-Хензела, Cochran–Mantel–Haenszel test)

G-тест

G-тест

G-тест основан на вычислении правдоподобий, но может применяться как тест для таблиц сопряженности.

H_0 : — категориальные переменные независимы друг от друга

H_A : — категориальные переменные зависимы

$$G = 2 \sum_{row=1}^r \sum_{col=1}^c Observed_{(row,col)} \cdot \ln \left(\frac{Observed_{(row,col)}}{Expected_{(row,col)}} \right)$$

При H_0 распределение $G \sim \chi^2$
с числом степеней свободы $df = (r - 1)(c - 1)$

Односторонний тест

Особенности G-теста

- При малых объемах выборок менее точен, чем другие (Agresti, 2002). Лучше использовать точный критерий Фишера.
- Подходит для данных со множеством переменных-предикторов (Sokal, Rholf, 1995; Agresti, 2002)

Summary

Summary

Таблицы сопряженности (contingency tables) показывают, как частоты категорий по одной переменной зависят от значения другой категориальной переменной.

Анализ сопряженности позволяет по таблице частот оценить, насколько связаны (“сопряжены”) друг с другом категориальные переменные.

Summary

- Риск - вероятность определенного исхода.
- Шансы - вероятность наступления определенного исхода, делённая на вероятность его не наступления
- Для описания связи между переменными используются
 - относительный риск - отношение выборочных оценок вероятностей исхода в сравниваемых группах
 - отношение шансов
- В исследованиях типа случай-контроль нельзя рассчитывать относительный риск, т.к. вероятности некоторых категорий заданы исследователем, а не свойствами выборки.

Summary

- Для анализа таблиц сопряженности используется тест сопряженности χ^2 , который позволяет протестировать гипотезу о независимости двух категориальных переменных.
- При нарушении условий применимости χ^2 -теста для тестиования гипотезы о независимости используют точный критерий Фишера.
- Для тестиования гипотезы о независимости категориальных переменных часто используют G-тест, но он плохо работает на малых выборках, по-этому предпочтительнее использовать точный критерий Фишера или обычный тест сопряженности χ^2 .

ЧТО ПОЧИТАТЬ

Agresti, A., Franklin, C. A., & Klingenberg, B. (2017). Statistics: The art and science of learning from data (Fourth edition). Pearson.

Whitlock, M., & Schluter, D. (2015). The analysis of biological data (Second edition). Roberts and Company Publishers.