

Пропорции и частоты

Основы биостатистики, осень 2022

Марина Варфоломеева

Пропорции и относительные частоты (доли)

- Биномиальное распределение
- Формула биномиального распределения
- Выборочное распределение доли
- Биномиальный тест для долей
- Доверительные интервалы для долей

Биномиальное распределение

Биномиальное распределение

Случайные события с двумя исходами:

- Бросок монетки:
 - орел
 - решка
- Использование рук
 - правша
 - левша
- Заболевание
 - заболел
 - остался здоров
- Сложная операция, тяжелая болезнь или катастрофа
 - выжил
 - умер

Условно одно из событий в паре называют “успех”, а другое — “неудача”.

Биномиальное распределение описывает вероятность того, сколько раз будет наблюдаться “успех” при определенном числе испытаний, когда вероятность успеха одинакова во всех испытаниях.

Следует ли X биномиальному распределению? (1)

У вас есть “нечестная” монетка. Вероятность, что выпадет орёл — 80%. Вы подбрасываете монетку совершенно одинаковым способом ровно 10 раз.

X — число выпавших орлов

Является ли X биномиально распределенной величиной?

Следует ли X биномиальному распределению? (1)

У вас есть “нечестная” монетка. Вероятность, что выпадет орёл — 80%. Вы подбрасываете монетку совершенно одинаковым способом ровно 10 раз.

X — число выпавших орлов

Является ли X биномиально распределенной величиной?

Да, потому что:

- В результате каждого броска монетки происходит выпадает орел или решка ([два события: “успех” или “неудача”](#)).

Следует ли X биномиальному распределению? (1)

У вас есть “нечестная” монетка. Вероятность, что выпадет орёл — 80%. Вы подбрасываете монетку совершенно одинаковым способом ровно 10 раз.

X — число выпавших орлов

Является ли X биномиально распределенной величиной?

Да, потому что:

- В результате каждого броска монетки происходит выпадает орел или решка ([два события: “успех” или “неудача”](#)).
- Монетку подбрасывают 10 раз совершенно одинаково ([испытания проходят одинаково](#)).

Следует ли X биномиальному распределению? (1)

У вас есть “нечестная” монетка. Вероятность, что выпадет орёл — 80%. Вы подбрасываете монетку совершенно одинаковым способом ровно 10 раз.

X — число выпавших орлов

Является ли X биномиально распределенной величиной?

Да, потому что:

- В результате каждого броска монетки происходит выпадает орел или решка ([два события: “успех” или “неудача”](#)).
- Монетку подбрасывают 10 раз совершенно одинаково ([испытания проходят одинаково](#)).
- Результат броска не влияет на результат других бросков ([испытания независимы друг от друга](#)).

Следует ли X биномиальному распределению? (1)

У вас есть “нечестная” монетка. Вероятность, что выпадет орёл — 80%. Вы подбрасываете монетку совершенно одинаковым способом ровно 10 раз.

X — число выпавших орлов

Является ли X биномиально распределенной величиной?

Да, потому что:

- В результате каждого броска монетки происходит выпадает орел или решка (**два события: “успех” или “неудача”**).
- Монетку подбрасывают 10 раз совершенно одинаково (**испытания проходят одинаково**).
- Результат броска не влияет на результат других бросков (**испытания независимы друг от друга**).
- Вероятность выпадения орла одинакова во всех бросках (**вероятность “успеха” одинакова**).

Следует ли X биномиальному распределению? (1)

У вас есть “нечестная” монетка. Вероятность, что выпадет орёл — 80%. Вы подбрасываете монетку совершенно одинаковым способом ровно 10 раз.

X — число выпавших орлов

Является ли X биномиально распределенной величиной?

Да, потому что:

- В результате каждого броска монетки происходит выпадает орел или решка (**два события: “успех” или “неудача”**).
- Монетку подбрасывают 10 раз совершенно одинаково (**испытания проходят одинаково**).
- Результат броска не влияет на результат других бросков (**испытания независимы друг от друга**).
- Вероятность выпадения орла одинакова во всех бросках (**вероятность “успеха” одинакова**).
- X — число орлов (**число “успехов”**)

Следует ли X биномиальному распределению? (1)

У вас есть “нечестная” монетка. Вероятность, что выпадет орёл — 80%. Вы подбрасываете монетку совершенно одинаковым способом ровно 10 раз.

X — число выпавших орлов

Является ли X биномиально распределенной величиной?

Да, потому что:

- В результате каждого броска монетки происходит выпадает орел или решка (**два события: “успех” или “неудача”**).
- Монетку подбрасывают 10 раз совершенно одинаково (**испытания проходят одинаково**).
- Результат броска не влияет на результат других бросков (**испытания независимы друг от друга**).
- Вероятность выпадения орла одинакова во всех бросках (**вероятность “успеха” одинакова**).
- X — число орлов (**число “успехов”**)

В этом примере перечислены все основные признаки биномиально-распределенной случайной величины.

Следует ли X биномиальному распределению? (2)

Университетский администратор обзванивает случайным образом выбранных выпускников до тех пор пока не найдет 5 человек, которые работают по специальности.

X — число выпускников, которых пришлось обзвонить

Является ли X биномиально распределенной величиной?

Следует ли X биномиальному распределению? (2)

Университетский администратор обзванивает случайным образом выбранных выпускников до тех пор пока не найдет 5 человек, которые работают по специальности.

X — число выпускников, которых пришлось обзвонить

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Общее число звонков (испытаний) не было заранее зафиксировано.
- X не равен числу выпускников, работающих по специальности.

Следует ли X биномиальному распределению? (3)

В пенале лежит 10 ручек. Человек последовательно выкладывает на стол случайные 4 из них.

X — число ручек, которые пишут, из всех выложенных на стол.

Является ли X биномимально распределенной величиной?

Следует ли X биномиальному распределению? (3)

В пенале лежит 10 ручек. Человек последовательно выкладывает на стол случайные 4 из них.

X — число ручек, которые пишут, из всех выложенных на стол.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность выбрать пишущую ручку не постоянна.

Следует ли X биномиальному распределению? (3)

В пенале лежит 10 ручек. Человек последовательно выкладывает на стол случайные 4 из них.

X — число ручек, которые пишут, из всех выложенных на стол.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность выбрать пишущую ручку не постоянна.

Допустим 6 из 10 ручек пишут.

Следует ли X биномиальному распределению? (3)

В пенале лежит 10 ручек. Человек последовательно выкладывает на стол случайные 4 из них.

X — число ручек, которые пишут, из всех выложенных на стол.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность выбрать пишущую ручку не постоянна.

Допустим 6 из 10 ручек пишут.

Тогда для 1й ручки $P(1\text{я ручка пишет}) = 6/10$

Следует ли X биномиальному распределению? (3)

В пенале лежит 10 ручек. Человек последовательно выкладывает на стол случайные 4 из них.

X — число ручек, которые пишут, из всех выложенных на стол.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность выбрать пишущую ручку не постоянна.

Допустим 6 из 10 ручек пишут.

Тогда для 1й ручки $P(1\text{я ручка пишет}) = 6/10$

Для 2й ручки возможно два варианта:

Следует ли X биномиальному распределению? (3)

В пенале лежит 10 ручек. Человек последовательно выкладывает на стол случайные 4 из них.

X — число ручек, которые пишут, из всех выложенных на стол.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность выбрать пишущую ручку не постоянна.

Допустим 6 из 10 ручек пишут.

Тогда для 1й ручки $P(1\text{я ручка пишет}) = 6/10$

Для 2й ручки возможно два варианта:

- $P(2\text{я ручка пишет} \mid 1\text{я пишет}) = 5/9$

Следует ли X биномиальному распределению? (3)

В пенале лежит 10 ручек. Человек последовательно выкладывает на стол случайные 4 из них.

X — число ручек, которые пишут, из всех выложенных на стол.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность выбрать пишущую ручку не постоянна.

Допустим 6 из 10 ручек пишут.

Тогда для 1й ручки $P(1\text{я ручка пишет}) = 6/10$

Для 2й ручки возможно два варианта:

- $P(2\text{я ручка пишет} \mid 1\text{я пишет}) = 5/9$
- $P(2\text{я ручка пишет} \mid 1\text{я не пишет}) = 6/9.$

Следует ли X биномиальному распределению? (3)

В пенале лежит 10 ручек. Человек последовательно выкладывает на стол случайные 4 из них.

X — число ручек, которые пишут, из всех выложенных на стол.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность выбрать пишущую ручку не постоянна.

Допустим 6 из 10 ручек пишут.

Тогда для 1й ручки $P(1\text{я ручка пишет}) = 6/10$

Для 2й ручки возможно два варианта:

- $P(2\text{я ручка пишет} \mid 1\text{я пишет}) = 5/9$
- $P(2\text{я ручка пишет} \mid 1\text{я не пишет}) = 6/9.$

Такая величина X подчиняется гипергеометрическому распределению.

Следует ли X биномиальному распределению? (4)

В университете опросили случайную выборку 1000 студентов.

X — число студентов из этой выборки, которые добираются до университета на велосипеде.

Является ли X биномиально распределенной величиной?

Следует ли X биномиальному распределению? (4)

В университете опросили случайную выборку 1000 студентов.

X — число студентов из этой выборки, которые добираются до университета на велосипеде.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность того, что студент велосипедист тоже не постоянна. X подчиняется гипергеометрическому распределению.

Следует ли X биномиальному распределению? (4)

В университете опросили случайную выборку 1000 студентов.

X — число студентов из этой выборки, которые добираются до университета на велосипеде.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность того, что студент велосипедист тоже не постоянна. X подчиняется гипергеометрическому распределению.

Допустим всего 10000 студентов и 1000 из них велосипедисты, $P(\text{велосипедист}) = 0.1$.

Следует ли X биномиальному распределению? (4)

В университете опросили случайную выборку 1000 студентов.

X — число студентов из этой выборки, которые добираются до университета на велосипеде.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность того, что студент велосипедист тоже не постоянна. X подчиняется гипергеометрическому распределению.

Допустим всего 10000 студентов и 1000 из них велосипедисты, $P(\text{велосипедист}) = 0.1$.

$$P(\text{1й велосипедист}) = 1000/10000 = 0.1$$

Следует ли X биномиальному распределению? (4)

В университете опросили случайную выборку 1000 студентов.

X — число студентов из этой выборки, которые добираются до университета на велосипеде.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность того, что студент велосипедист тоже не постоянна. X подчиняется гипергеометрическому распределению.

Допустим всего 10000 студентов и 1000 из них велосипедисты, $P(\text{велосипедист}) = 0.1$.

$$P(\text{1-й велосипедист}) = 1000/10000 = 0.1$$

$$P(\text{2-й велосипедист} \mid \text{1-й велосипедист}) = 999/9999 = 0.099909991$$

Следует ли X биномиальному распределению? (4)

В университете опросили случайную выборку 1000 студентов.

X — число студентов из этой выборки, которые добираются до университета на велосипеде.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность того, что студент велосипедист тоже не постоянна. X подчиняется гипергеометрическому распределению.

Допустим всего 10000 студентов и 1000 из них велосипедисты, $P(\text{велосипедист}) = 0.1$.

$$P(1\text{й велосипедист}) = 1000/10000 = 0.1$$

$$P(2\text{й велосипедист} \mid 1\text{й велосипедист}) = 999/9999 = 0.099909991$$

$$P(2\text{й велосипедист} \mid 1\text{й пешеход}) = 1000/9999 = 0.100010001$$

и т.д.

Следует ли X биномиальному распределению? (4)

В университете опросили случайную выборку 1000 студентов.

X — число студентов из этой выборки, которые добираются до университета на велосипеде.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность того, что студент велосипедист тоже не постоянна. X подчиняется гипергеометрическому распределению.

Допустим всего 10000 студентов и 1000 из них велосипедисты, $P(\text{велосипедист}) = 0.1$.

$$P(1\text{й велосипедист}) = 1000/10000 = 0.1$$

$$P(2\text{й велосипедист} \mid 1\text{й велосипедист}) = 999/9999 = 0.099909991$$

$$P(2\text{й велосипедист} \mid 1\text{й пешеход}) = 1000/9999 = 0.100010001$$

и т.д.

Однако эти величины мало отличаются от 0.1.

Следует ли X биномиальному распределению? (4)

В университете опросили случайную выборку 1000 студентов.

X — число студентов из этой выборки, которые добираются до университета на велосипеде.

Является ли X биномиально распределенной величиной?

Нет, потому что:

- Вероятность того, что студент велосипедист тоже не постоянна. X подчиняется гипергеометрическому распределению.

Допустим всего 10000 студентов и 1000 из них велосипедисты, $P(\text{велосипедист}) = 0.1$.

$$P(1\text{й велосипедист}) = 1000/10000 = 0.1$$

$$P(2\text{й велосипедист} \mid 1\text{й велосипедист}) = 999/9999 = 0.099909991$$

$$P(2\text{й велосипедист} \mid 1\text{й пешеход}) = 1000/9999 = 0.100010001$$

и т.д.

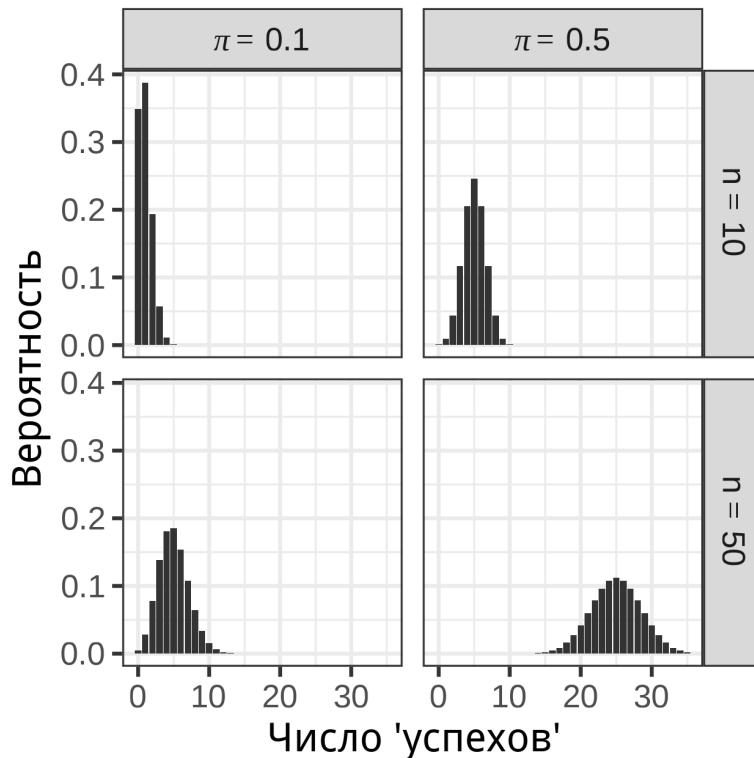
Однако эти величины мало отличаются от 0.1.

Если объем выборки n мал в сравнении с объемом генеральной совокупности N , гипергеометрическое распределение можно аппроксимировать биномиальным.

Формула биномиального распределения

Формула биномиального распределения

$$P(X \text{ 'успехов'}) = \binom{n}{X} \pi^X (1 - \pi)^{n-X} = \frac{n!}{X!(n - X)!} \pi^X (1 - \pi)^{n-X}$$



Параметры:

- n — число испытаний (фиксированное), испытания независимы друг от друга
- π — вероятность успеха одинакова во всех испытаниях.

Биномиальный коэффициент описывает
число сочетаний из n по X

$$\binom{n}{X} = C_n^X = \frac{n!}{X! \cdot (n - X)!}$$

Смысл биномиального коэффициента (1)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

Смысл биномиального коэффициента (1)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

1-я зараженная улитка может быть расположена 5-ю способами. Допустим, так:

(-) **(i-1)** (-) (-) (-)

Смысл биномиального коэффициента (1)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

1-я зараженная улитка может быть расположена 5-ю способами. Допустим, так:

(-) **(i-1)** (-) (-) (-)

2-я — 4-мя способами, допустим так:

(-) **(i-1)** (-) (-) **(i-2)**

Смысл биномиального коэффициента (1)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

1-я зараженная улитка может быть расположена 5-ю способами. Допустим, так:

(-) (i-1) (-) (-) (-)

2-я — 4-мя способами, допустим так:

(-) (i-1) (-) (-) (i-2)

Третья — 3-мя способами, допустим так:

(i-3) (i-1) (-) (-) (i-2)

Общее число способов, которыми могут располагаться эти зараженные улитки, если бы они были различны: $5 \cdot 4 \cdot 3$

Смысл биномиального коэффициента (1)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

1-я зараженная улитка может быть расположена 5-ю способами. Допустим, так:

(-) (i-1) (-) (-) (-)

2-я — 4-мя способами, допустим так:

(-) (i-1) (-) (-) (i-2)

Третья — 3-мя способами, допустим так:

(i-3) (i-1) (-) (-) (i-2)

Общее число способов, которыми могут располагаться эти зараженные улитки, если бы они были различимы: $5 \cdot 4 \cdot 3$

Но нам не важен порядок, поэтому разные варианты расположения трех зараженных улиток эквивалентны. Например вот эти два (а всего их $3 \cdot 2 \cdot 1$):

(i-1) (i-2) (-) (-) (i-3)

(i-3) (i-2) (-) (-) (i-1)

Поэтому число способов, которыми могут располагаться зараженные улитки, если они не различимы (т.е. если нам не важен их порядок):

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10$$

Смысл биномиального коэффициента (2)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

Смысл биномиального коэффициента (2)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

1-я здоровая улитка может быть расположена 5-ю способами. Допустим, так:

(-) (-) (-) (**h-1**) (-)

Смысл биномиального коэффициента (2)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

1-я здоровая улитка может быть расположена 5-ю способами. Допустим, так:

(-) (-) (-) (**h-1**) (-)

2-я здоровая — 4-мя способами, допустим так:

(**h-2**) (-) (-) (**h-1**) (-)

Общее число способов, которыми могут располагаться эти 2 здоровые улитки, если бы они были различны: $5 \cdot 4$

Смысл биномиального коэффициента (2)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

(-) (-) (-) (-) (-)

1-я здоровая улитка может быть расположена 5-ю способами. Допустим, так:

(-) (-) (-) (**h-1**) (-)

2-я здоровая — 4-мя способами, допустим так:

(**h-2**) (-) (-) (**h-1**) (-)

Общее число способов, которыми могут располагаться эти 2 здоровые улитки, если бы они были различимы: $5 \cdot 4$

Но нам не важен порядок, поэтому разные варианты расположения 2 здоровых улиток эквивалентны (и всего их $2 \cdot 1$):

(**h-2**) (-) (-) (**h-1**) (-)

(**h-1**) (-) (-) (**h-2**) (-)

Поэтому число способов, которыми могут располагаться 2 здоровые улитки, если они не различимы (т.е. если нам не важен их порядок):

$\frac{5 \cdot 4}{2 \cdot 1} = 10$, так же, как в прошлый раз, но формула другая. Непорядок!

Смысл биномиального коэффициента (3)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

Как из этих двух формул сделать одну, чтобы было все равно с кого начинать — со здоровы или зараженных улиток?

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10, \frac{5 \cdot 4}{2 \cdot 1} = 10$$

Смысл биномиального коэффициента (3)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

Как из этих двух формул сделать одну, чтобы было все равно с кого начинать — со здоровы или зараженных улиток?

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10, \quad \frac{5 \cdot 4}{2 \cdot 1} = 10$$

Домножим числитель и знаменатель на одно и то же число:

$$\frac{(5 \cdot 4 \cdot 3) \cdot (2 \cdot 1)}{(3 \cdot 2 \cdot 1) \cdot (2 \cdot 1)} = \frac{5!}{3!2!}$$

Смысл биномиального коэффициента (3)

Пусть перед нами 5 улиток: 3 **заражены трематодами (i)**, а 2 **здоровы (h)**. Сколькими возможными способами могут быть расположены эти улитки?

Как из этих двух формул сделать одну, чтобы было все равно с кого начинать — со здоровы или зараженных улиток?

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10, \quad \frac{5 \cdot 4}{2 \cdot 1} = 10$$

Домножим числитель и знаменатель на одно и то же число:

$$\frac{(5 \cdot 4 \cdot 3) \cdot (2 \cdot 1)}{(3 \cdot 2 \cdot 1) \cdot (2 \cdot 1)} = \frac{5!}{3!2!}$$

Легко заметить, что в числителе факториал общего числа улиток, а в знаменателе факториалы каждой из категорий.

$$\binom{n}{X} = C_n^X = \frac{n!}{X! \cdot (n - X)!}$$

Вычисляем вероятность с использованием биномиального распределения

Вы собираете улиток для эксперимента.

Вероятность того, что случайно выбранная улитка
окажется зараженной trematодами 0.2.

Какова вероятность того, что из 5 собранных улиток 3 заражены?

Вычисляем вероятность с использованием биномиального распределения

Вы собираете улиток для эксперимента.

Вероятность того, что случайно выбранная улитка окажется зараженной трематодами 0.2.

Какова вероятность того, что из 5 собранных улиток 3 заражены?

- $0.2^3(1 - 0.2)^2$ — вероятность того, что из 5 улиток 3 заражены, а 2 здоровы
- $\frac{5!}{3!2!}$ — число возможных сочетаний здоровых и зараженных улиток

Вычисляем вероятность с использованием биномиального распределения

Вы собираете улиток для эксперимента.

Вероятность того, что случайно выбранная улитка окажется зараженной трематодами 0.2.

Какова вероятность того, что из 5 собранных улиток 3 заражены?

- $0.2^3(1 - 0.2)^2$ — вероятность того, что из 5 улиток 3 заражены, а 2 здоровы
- $\frac{5!}{3!2!}$ — число возможных сочетаний здоровых и зараженных улиток

Можно воспользоваться биномиальным распределением:

$$P(X \text{ 'успехов'}) = \frac{n!}{X!(n-X)!} \pi^X (1 - \pi)^{n-X}$$

$$P(3) = \frac{5!}{3!2!} 0.2^3(1 - 0.2)^2 = 10 \cdot 0.008 \cdot 0.64 = 0.0512$$

Биномиальное распределение (количество “успехов”)

Пользуясь формулой биномиального распределения можно рассчитать p — вероятность получить любое количество успехов X в n испытаниях.

$$P(X \text{ 'успехов'}) = \frac{n!}{X!(n-X)!} \pi^X (1-\pi)^{n-X}$$

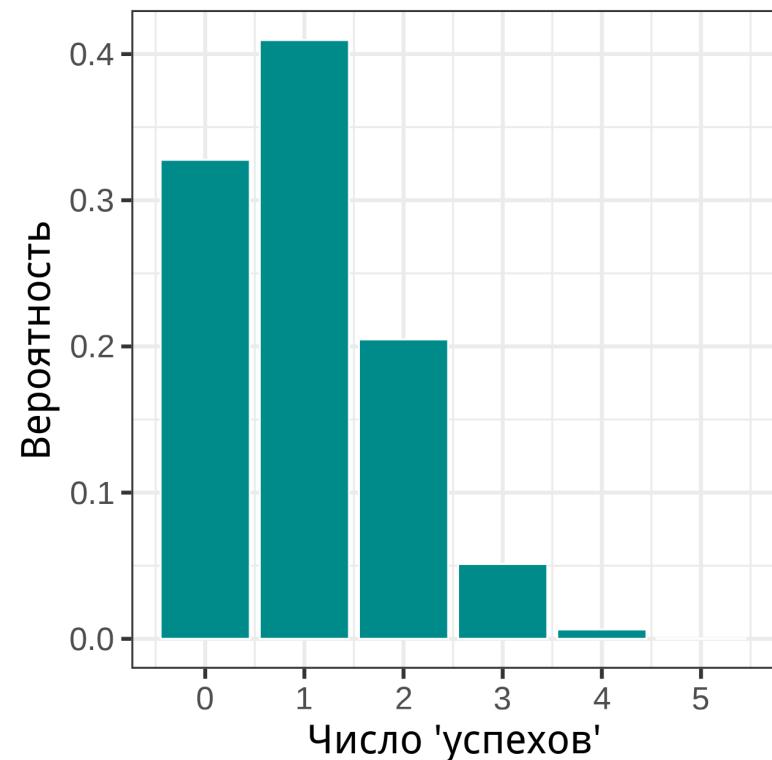
Биномиальное распределение (количество “успехов”)

Пользуясь формулой биномиального распределения можно рассчитать p — вероятность получить любое количество успехов X в n испытаниях.

$$P(X \text{ 'успехов'}) = \frac{n!}{X!(n-X)!} \pi^X (1-\pi)^{n-X}$$

Т.е. для нашего примера p — вероятность получить определенное количество зараженных моллюсков в выборке из 5 особей.

x	p
0	0.328
1	0.410
2	0.205
3	0.051
4	0.006
5	0.000



Выборочное распределение доли

Доля “успехов”

Для любой выборки можно оценить p — долю “успехов” X среди всех n испытаний:

$$p = \frac{X}{n}$$

Доля “успехов”

Для любой выборки можно оценить p — долю “успехов” X среди всех n испытаний:

$$p = \frac{X}{n}$$

В примере с улитками была выборка из 5 улиток, 3 из которых были заражены.

Доля зараженных улиток $p = \frac{3}{5} = 0.6$

В другой выборке была бы другая доля.

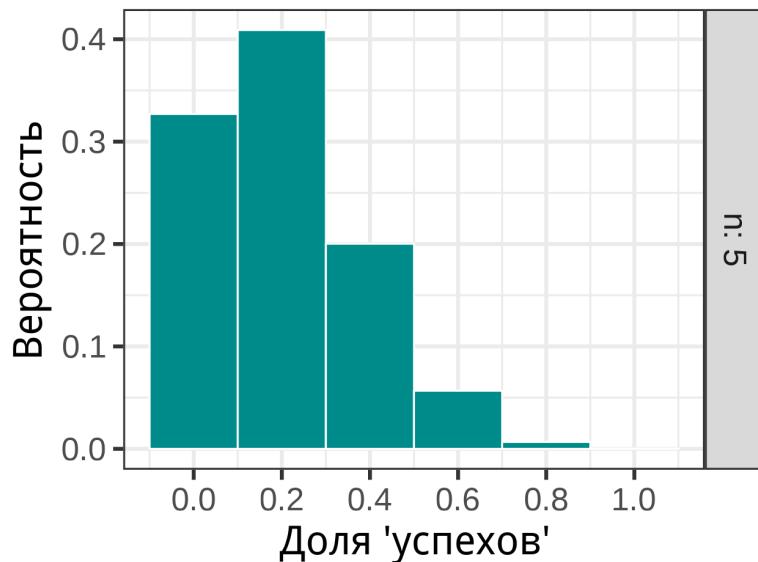
Выборочное распределение доли

Если взять множество выборок размером n и в каждой оценить долю “успехов”, то получится **выборочное распределение доли**.

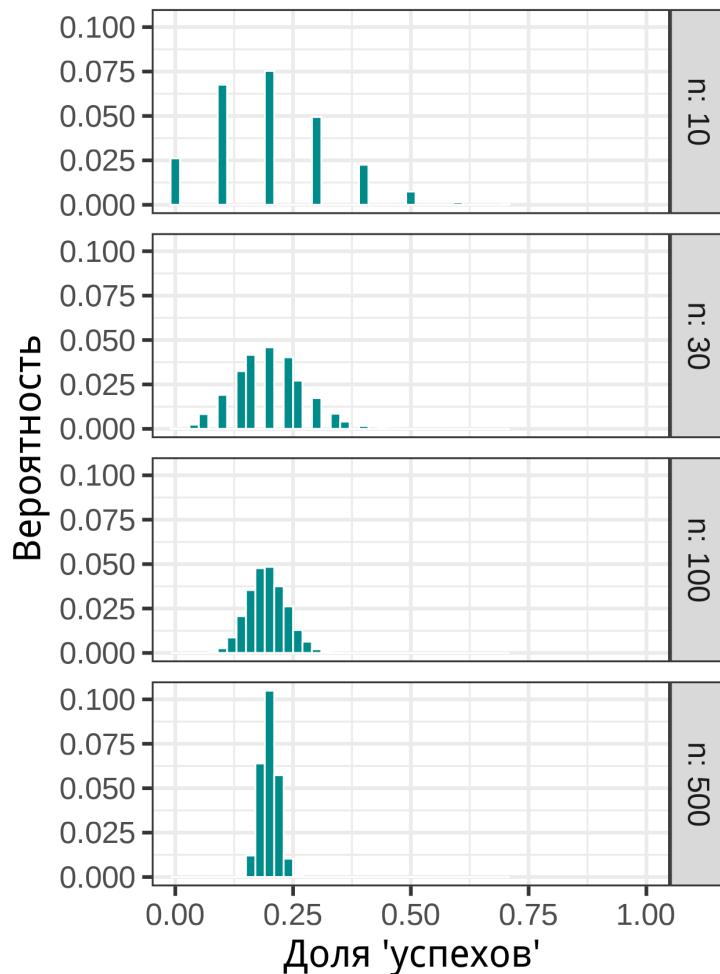
Выборочное распределение доли

Если взять множество выборок размером n и в каждой оценить долю “успехов”, то получится **выборочное распределение доли**.

Например, вот распределение доли зараженных моллюсков из нашего примера (выборки по 5 улиток, при вероятности заражения 0.2).



Параметры выборочного распределения доли



Форма выборочного распределения доли зависит от объема выборки

Параметры выборочного распределения доли:

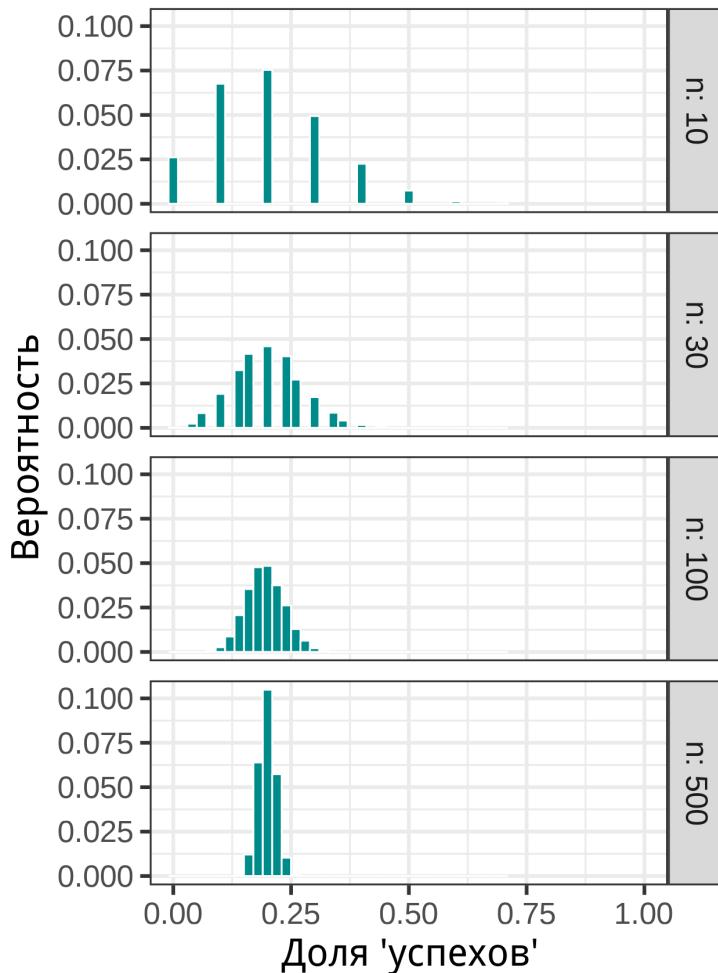
- **среднее:**

$$\mu_p = \pi$$

- **стандартное отклонение**
(стандартная ошибка доли):

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Параметры выборочного распределения доли



Форма выборочного распределения доли зависит от объема выборки

Параметры выборочного распределения доли:

- **среднее:**

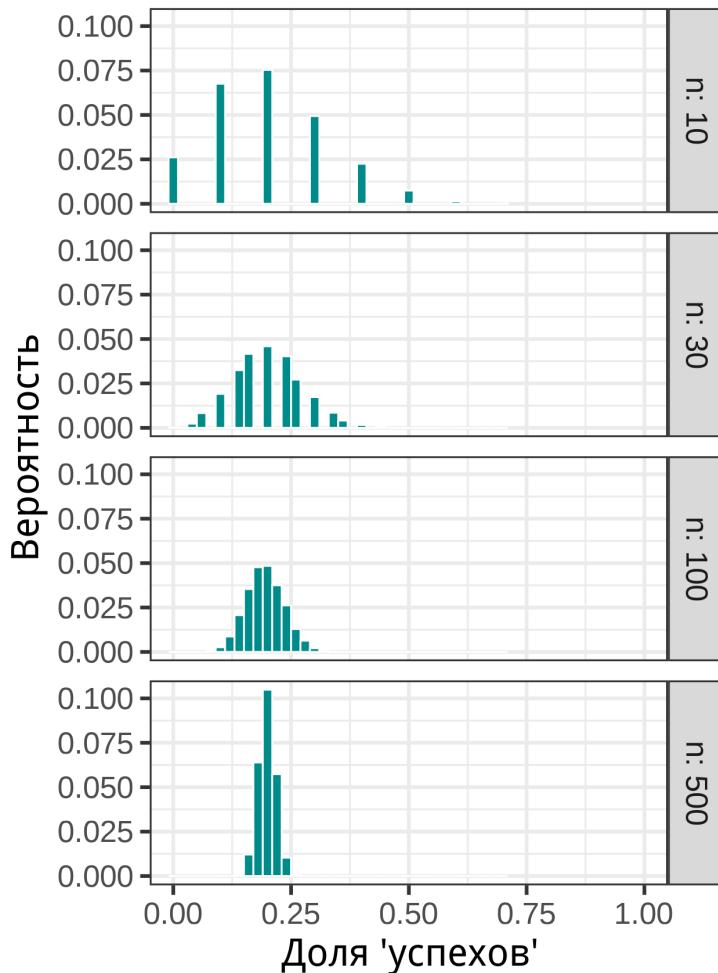
$$\mu_p = \pi$$

- **стандартное отклонение**
(стандартная ошибка доли):

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Т.е. точность оценки доли по выборке будет зависеть от объема выборки.

Параметры выборочного распределения доли



Форма выборочного распределения доли зависит от объема выборки

Параметры выборочного распределения доли:

- **среднее:**

$$\mu_p = \pi$$

- **стандартное отклонение**
(стандартная ошибка доли):

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Стандартную ошибку оценивают по выборке:

$$SE_p = \sqrt{\frac{p(1 - p)}{n}}$$

Выборочное распределение доли используют для построения доверительных интервалов к долям.

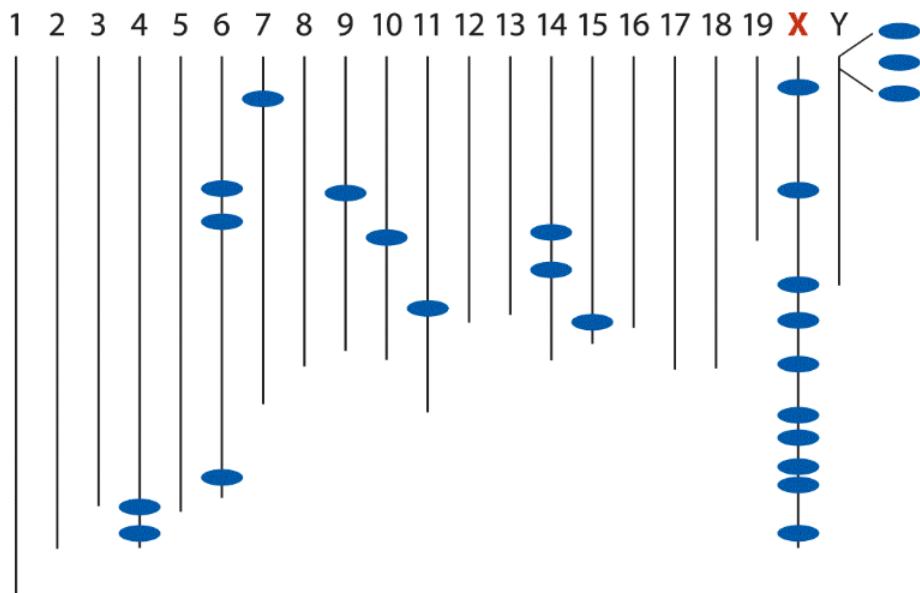
Биномиальный тест для долей

Пример: гены сперматогенеза на X хромосоме мышей

Из 25 генов сперматогенеза у мышей 10 (40%) расположены на X хромосоме (Wang et al. 2001).

X хромосома мышей содержит 6.1% всех генов.

Поэтому, если бы гены были распределены по хромосомам случайно, только 6.1% были бы на X хромосоме.



Свидетельствуют ли эти данные, что гены сперматогенеза у мышей непропорционально часто встречаются на X хромосоме?

рис. 7.2 из Whitlock and Schluter, 2015

Биномиальный тест для долей

Биномиальный тест использует биномиальное распределение, чтобы проверить значимость отличия доли “успехов” в серии из n испытаний от какой-то величины.

$$P(X \text{ 'успехов'}) = \frac{n!}{X!(n-X)!} \pi^X (1 - \pi)^{n-X}$$

$H_0 : \pi = p_0$ — доля “успехов” в генеральной совокупности равна p_0

$H_A : \pi \neq p_0$ — доля “успехов” в генеральной совокупности не равна p_0

$$0 \leq p_0 \leq 1$$

Тестовая статистика — наблюдаемое число “успехов”

Биномиальный тест для долей

Биномиальный тест использует биномиальное распределение, чтобы проверить значимость отличия доли “успехов” в серии из n испытаний от какой-то величины.

$$P(X \text{ 'успехов'}) = \frac{n!}{X!(n-X)!} \pi^X (1-\pi)^{n-X}$$

$H_0 : \pi = p_0$ — доля “успехов” в генеральной совокупности равна p_0

$H_A : \pi \neq p_0$ — доля “успехов” в генеральной совокупности не равна p_0

$0 \leq p_0 \leq 1$

Тестовая статистика — наблюдаемое число “успехов”

В примере про гены мышей

Испытание: “попал” ген сперматогенеза на X хромосому (= “успех”) или нет.

Всего 25 генов (испытаний).

$H_0 : \pi = 0.061$ — на X хромосоме расположено 6.1% генов сперматогенеза

$H_A : \pi \neq 0.061$ — на X хромосоме другой процент генов сперматогенеза

Тестовая статистика — наблюдаемое число генов сперматогенеза на X хромосоме

Биномиальный тест для долей

Биномиальный тест использует биномиальное распределение, чтобы проверить значимость отличия доли “успехов” в серии из n испытаний от какой-то величины.

$$P(X \text{ 'успехов'}) = \frac{n!}{X!(n-X)!} \pi^X (1-\pi)^{n-X}$$

$H_0 : \pi = p_0$ — доля “успехов” в генеральной совокупности равна p_0

$H_A : \pi \neq p_0$ — доля “успехов” в генеральной совокупности не равна p_0

$0 \leq p_0 \leq 1$

Тестовая статистика — наблюдаемое число “успехов”

В примере про гены мышей

Испытание: “попал” ген сперматогенеза на X хромосому (= “успех”) или нет.
Всего 25 генов (испытаний).

$H_0 : \pi = 0.061$ — на X хромосоме расположено 6.1% генов сперматогенеза

$H_A : \pi \neq 0.061$ — на X хромосоме другой процент генов сперматогенеза

Тестовая статистика — наблюдаемое число генов сперматогенеза на X хромосоме

Как выглядит биномиальное распределение для ситуации, когда верна H_0 ?

Биномиальное распределение для H_0

$H_0 : \pi = 0.061$ — на X хромосоме расположено 6.1% генов сперматогенеза
 $H_A : \pi \neq 0.061$ — на X хромосоме другой процент генов сперматогенеза

Если H_0 справедлива, то число генов, попавших на X хромосому из 25 генов сперматогенеза описывается биномиальным распределением с параметрами $n = 25$ и $\pi = 0.061$

$$P(X \text{ генов сперматогенеза на X хр.}) = \frac{25!}{X!(25 - X)!} 0.061^X (1 - 0.061)^{25-X}$$

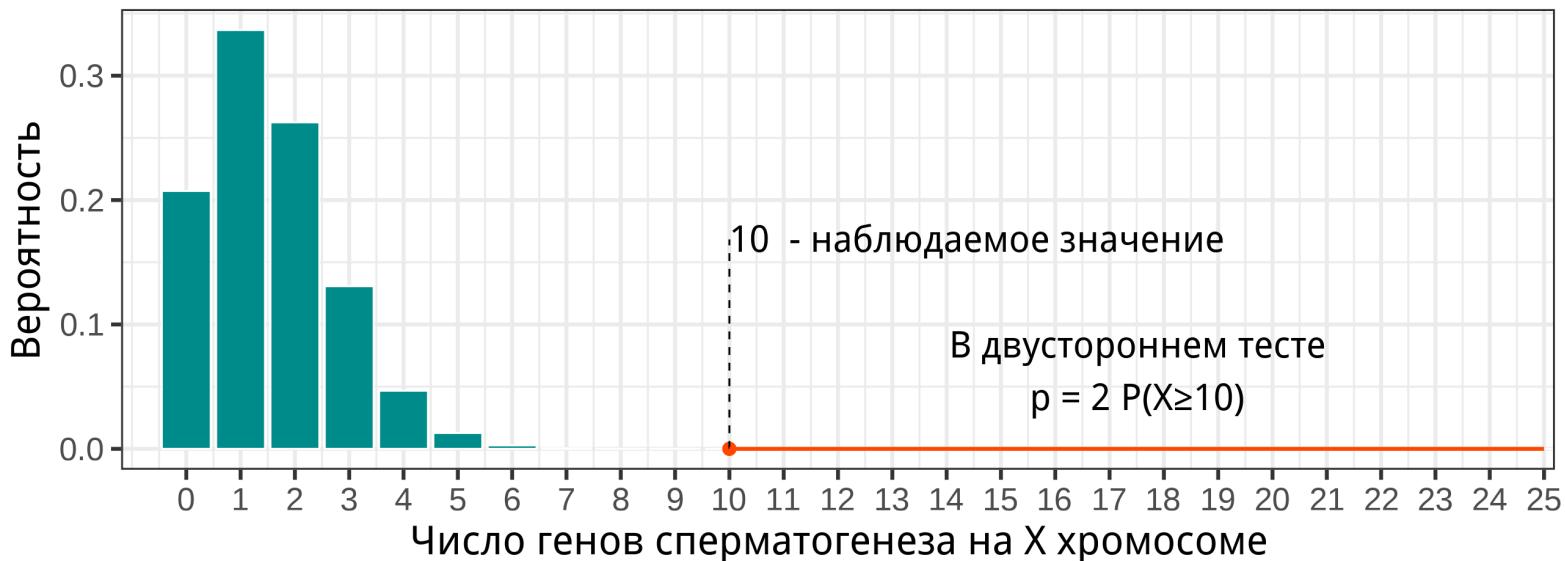


P-значение в двустороннем биномиальном teste

$H_0 : \pi = 0.061$ — на X хромосоме расположено 6.1% генов сперматогенеза
 $H_A : \pi \neq 0.061$ — на X хромосоме другой процент генов сперматогенеза

Наш тест двусторонний (см H_A), поэтому *p*-значение — это

- $P(x \geq 10)$ — вероятность получить при H_0 10 или более генов сперматогенеза на X хромосоме (правый хвост)
- плюс такая же вероятность получить слишком малое количество генов (на другом конце распределения)

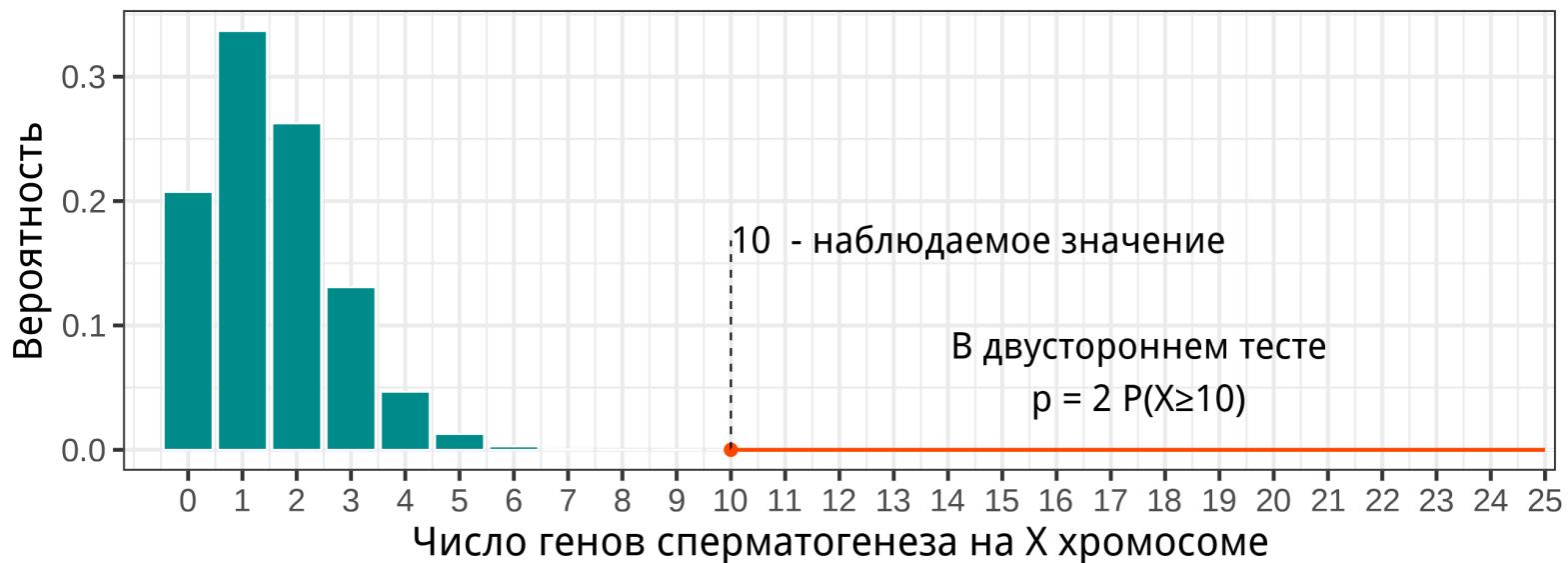


P-значение в двустороннем биномиальном teste

$H_0 : \pi = 0.061$ — на X хромосоме расположено 6.1% генов сперматогенеза
 $H_A : \pi \neq 0.061$ — на X хромосоме другой процент генов сперматогенеза

Чтобы вычислить р-значение, суммируем и удвоим вероятности $P(x \geq 10)$:

X	p	X	p	X	p	X	p
10	9.07e-07	14	2.2e-11	18	4.23e-17	22	3.61e-24
11	8.04e-08	15	1.05e-12	19	1.01e-18	23	3.06e-26
12	6.09e-09	16	4.26e-14	20	1.97e-20	24	1.65e-28
13	3.96e-10	17	1.47e-15	21	3.05e-22	25	4.3e-31

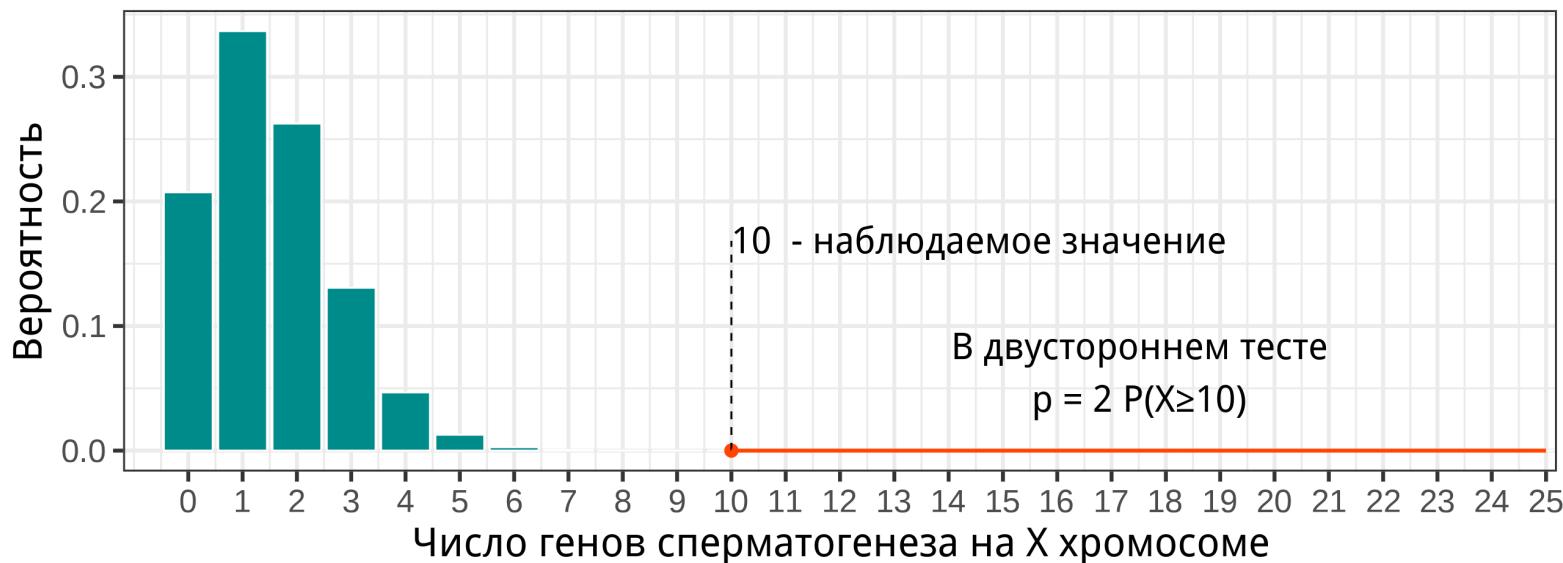


P-значение в двустороннем биномиальном тесте

$H_0 : \pi = 0.061$ — на X хромосоме расположено 6.1% генов сперматогенеза
 $H_A : \pi \neq 0.061$ — на X хромосоме другой процент генов сперматогенеза

Если суммировать и удвоить вероятности $P(x \geq 10)$ получится $1.99e-06 < 0.05$, т.е. H_0 придется отвергнуть.

На X хромосоме находится непропорционально большая доля генов сперматогенеза (0.40, SE = 0.10; биномиальный тест, n = 25, p < 0.001).



Апроксимации биномиального теста

Биномиальный тест дает точные значения p , но вычисления лучше делать не вручную.

В те времена, когда компьютеры были редкостью, разработали более быстрые приближенные варианты этого теста, которые вы можете встретить:

- χ^2 тест качества подгонки (χ^2 goodness of fit test)
- Нормальная аппроксимация биномиального теста

Доверительный интервал для долей

Оценка доли

$$p = \frac{X}{n}$$

Оценка доли

$$p = \frac{X}{n}$$

В нашем примере

$$p = \frac{10}{25} = 0.4$$

То есть 40% генов сперматогенеза у мышей лежат на X хромосоме.

Это точечная оценка. Давайте попробуем построить к ней доверительный интервал.

Стандартная ошибка доли

Когда обсуждали выборочное распределение доли, упоминали, что **стандартное отклонение выборочного распределения доли**:

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Ее выборочная оценка — это **стандартная ошибка доли** $SE_p = \sqrt{\frac{p(1 - p)}{n}}$

Стандартная ошибка доли

Когда обсуждали выборочное распределение доли, упоминали, что **стандартное отклонение выборочного распределения доли**:

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Ее выборочная оценка — это **стандартная ошибка доли** $SE_p = \sqrt{\frac{p(1 - p)}{n}}$

В нашем примере $SE_p = \sqrt{\frac{0.4(1 - 0.4)}{25}} = 0.098$

Эта величина характеризует степень точности нашей оценки доли генов на X хромосоме в генеральной совокупности.

Доверительный интервал для долей

Существует много методов построения приблизительных доверительных интервалов для долей.

Метод Вальда (Wald, 1939)

Предложен П.С. Лапласом в 1812г.

$$p \pm 1.96SE_p$$

Метод Вальда (Wald, 1939)

Предложен П.С. Лапласом в 1812г.

$$p \pm 1.96SE_p$$

Ограничения и недостатки:

- Плохо работает при малых объемах выборок
- Когда p близко к 0 или 1 получаются интервалы с выходом за пределы интервала [0, 1]
- Невозможен для $p = 0$ и 1
- Когда $np < 5$ или $n(1-p) < 5$ не правильная оценка SE_p т.к. не работает аппроксимация нормальным распределением (Motulsky, 1995)

Метод Вальда (Wald, 1939)

Предложен П.С. Лапласом в 1812г.

$$p \pm 1.96SE_p$$

Ограничения и недостатки:

- Плохо работает при малых объемах выборок
- Когда p близко к 0 или 1 получаются интервалы с выходом за пределы интервала $[0, 1]$
- Невозможен для $p = 0$ и 1
- Когда $np < 5$ или $n(1-p) < 5$ не правильная оценка SE_p т.к. не работает аппроксимация нормальным распределением (Motulsky, 1995)

В нашем примере $p = 0.4$, $SE_p = 0.098$, тогда предел погрешности $1.96 \cdot 0.098 = 0.192$.

Таким образом, на X хромосоме $40 \pm 19.2\%$ генов сперматогенеза (дов. инт. методом Вальда).

Метод Агрести-Коулл (Agresti, Coull, 1998)

Это модифицированный метод Вальда.

1. Считаем p с поправкой $p' = \frac{X + 2}{n + 4}$

2. Считаем стандартную ошибку доли с поправкой $SE'_p = \sqrt{\frac{p'(1 - p')}{n + 4}}$

3. Вычисляем границы доверительного интервала доли

$$p' - 1.96 \cdot SE'_p < p < p' + 1.96 \cdot SE'_p$$

Метод Агрести-Коулл предпочтительнее, чем метод Вальда.

Метод Агрести-Коулл (Agresti, Coull, 1998)

Это модифицированный метод Вальда.

1. Считаем p с поправкой $p' = \frac{X + 2}{n + 4}$

2. Считаем стандартную ошибку доли с поправкой $SE'_p = \sqrt{\frac{p'(1 - p')}{n + 4}}$

3. Вычисляем границы доверительного интервала доли

$$p' - 1.96 \cdot SE'_p < p < p' + 1.96 \cdot SE'_p$$

Метод Агрести-Коулл предпочтительнее, чем метод Вальда.

В нашем примере $n = 25$, $X = 10$, значит $p' = \frac{10 + 2}{25 + 4} = 0.414$.

Стандартная ошибка с поправкой $SE'_p = \sqrt{\frac{0.414(1 - 0.414)}{25 + 4}} = 0.0915$ тогда предел погрешности $1.96 \cdot 0.0915 = 0.179$.

Таким образом, на X хромосоме $40 \pm 17.9\%$ генов сперматогенеза (дов. инт. методом Агрести-Коулл).

Метод Вилсона (Wilson, 1927)

$$\text{Н. гр.} = \frac{p + \frac{z_{\text{H.}}^2}{2n} + z_{\text{H.}} \sqrt{\frac{p(1-p)}{n} + \frac{z_{\text{H.}}^2}{4n^2}}}{1 + \frac{z_{\text{H.}}^2}{n}}$$

$$\text{В. гр.} = \frac{p + \frac{z_{\text{B.}}^2}{2n} + z_{\text{B.}} \sqrt{\frac{p(1-p)}{n} + \frac{z_{\text{B.}}^2}{4n^2}}}{1 + \frac{z_{\text{B.}}^2}{n}}$$

Для 95% доверительного интервала $z_{\text{H.}} = -1.96$, $z_{\text{B.}} = 1.96$

Особенности:

- Не сильно зависит от n или p
- Асимметричный
- Нижняя граница доверительного интервала не бывает отрицательной (!)

Рекомендован к использованию (Agresti, Coull, 1998).

Метод Вилсона (Wilson, 1927)

$$\text{Н. гр.} = \frac{p + \frac{z_{\text{H.}}^2}{2n} + z_{\text{H.}} \sqrt{\frac{p(1-p)}{n} + \frac{z_{\text{H.}}^2}{4n^2}}}{1 + \frac{z_{\text{H.}}^2}{n}}$$

$$\text{В. гр.} = \frac{p + \frac{z_{\text{B.}}^2}{2n} + z_{\text{B.}} \sqrt{\frac{p(1-p)}{n} + \frac{z_{\text{B.}}^2}{4n^2}}}{1 + \frac{z_{\text{B.}}^2}{n}}$$

Для 95% доверительного интервала $z_{\text{H.}} = -1.96$, $z_{\text{B.}} = 1.96$

Особенности:

- Не сильно зависит от n или p
- Асимметричный
- Нижняя граница доверительного интервала не бывает отрицательной (!)

Рекомендован к использованию (Agresti, Coull, 1998).

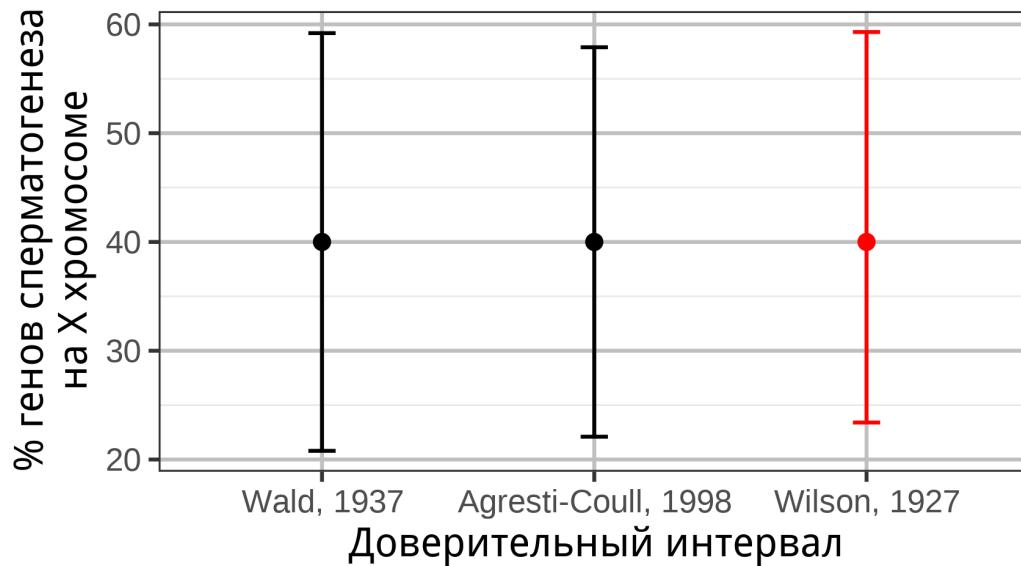
В нашем примере $n = 25$, $p = 0.4$, $\frac{p(1-p)}{n} = 0.0096$, значит

$$\text{н.гр.} = \frac{0.4 + \frac{1.96^2}{2.25} + 1.96 \sqrt{0.0096 + \frac{1.96^2}{4.25^2}}}{1 + \frac{1.96^2}{25}} = 0.234$$

$$\text{в.гр.} = \frac{0.4 + \frac{(-1.96)^2}{2.25} + (-1.96) \sqrt{0.0096 + \frac{(-1.96)^2}{4.25^2}}}{1 + \frac{(-1.96)^2}{25}} = 0.593$$

Т. обр., на X хромосоме 40% генов
сперматогенеза (дов. инт. методом Вилсона [23.4, 59.3] %).

Сравнение доверительных интервалов, полученных разными методами



Summary

Summary

Биномиальное распределение описывает вероятность того, сколько раз будет наблюдаться “успех” при определенном числе испытаний, когда вероятность успеха одинакова во всех испытаниях.

В формуле биномиального распределения биномиальный коэффициент описывает число возможных сочетаний заданного числа “успехов” и “неудач” при данном числе испытаний.

Summary

Выборочное распределение доли описывает, чему будет равна доля “успехов” в повторных выборках заданного размера при заданной вероятности “успеха”.

Стандартное отклонение выборочного распределения доли называют стандартной ошибкой доли. Она описывает точность оценки доли по данным выборки.

Биномиальный тест позволяет при помощи биномиального распределения тестировать гипотезы о том, что доля “успехов” в серии испытаний равна какой-то величине.

Summary

Доверительные интервалы к доле можно получить разными способами.

Метод Вальда самый простой в расчетах, но дает невозможные значения для очень малых и очень больших долей.

Метод Агрести-Коул лучше, чем метод Вальда, потому что включает поправку.

Метод Вилсона — это рекомендуемый метод построения доверительного интервала — дает асимметричные интервалы, не выходящие за пределы возможных значений долей.

ЧТО ПОЧИТАТЬ

Agresti, A., Franklin, C. A., & Klingenberg, B. (2017). Statistics: The art and science of learning from data (Fourth edition). Pearson. — глава **6.3 Probabilities When Each Observation Has Two Possible Outcomes**

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