Доверительные интервалы

Миленькин Александр
Биоинформатик в Insilico Medicine

Александр Миленькин

Биоинформатик в Insilico Medicine

О спикере:

- Преподаю в Нетологии
- Окончил МФТИ в 2019 году
- Активно участвую в соревнованиях по Data Science.

Аккаунты в соц.сетях

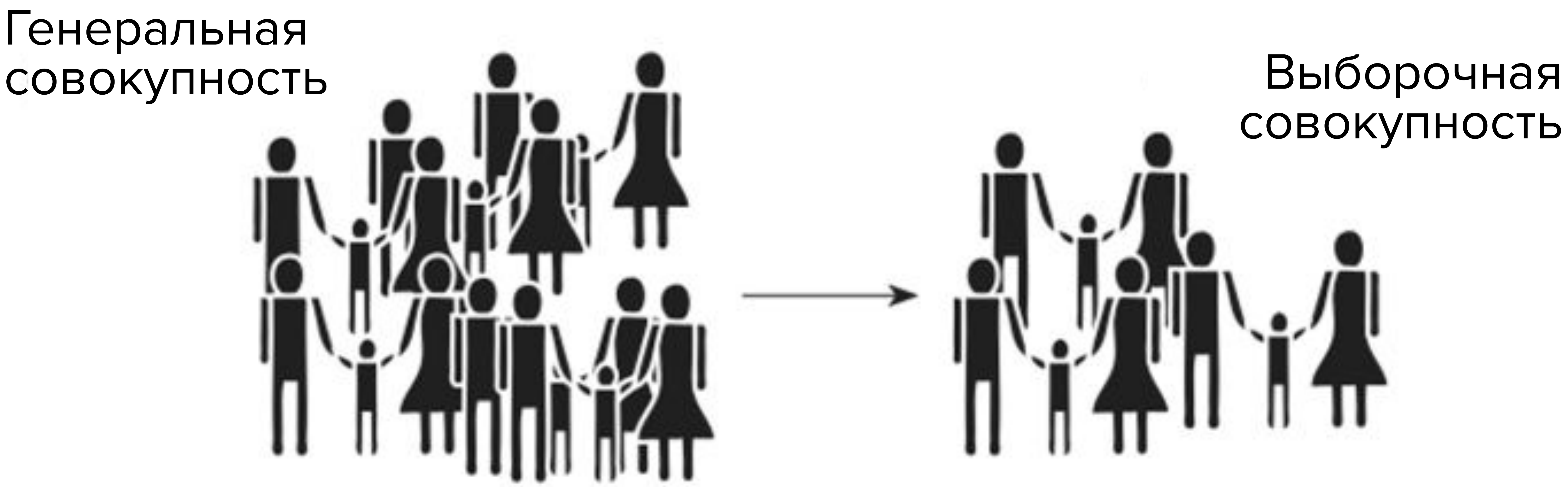
@Aleron75infskin

План урока

1. Доверительные интервалы
2. Статистическая проверка гипотез для несвязанных выборок

Точечная оценка

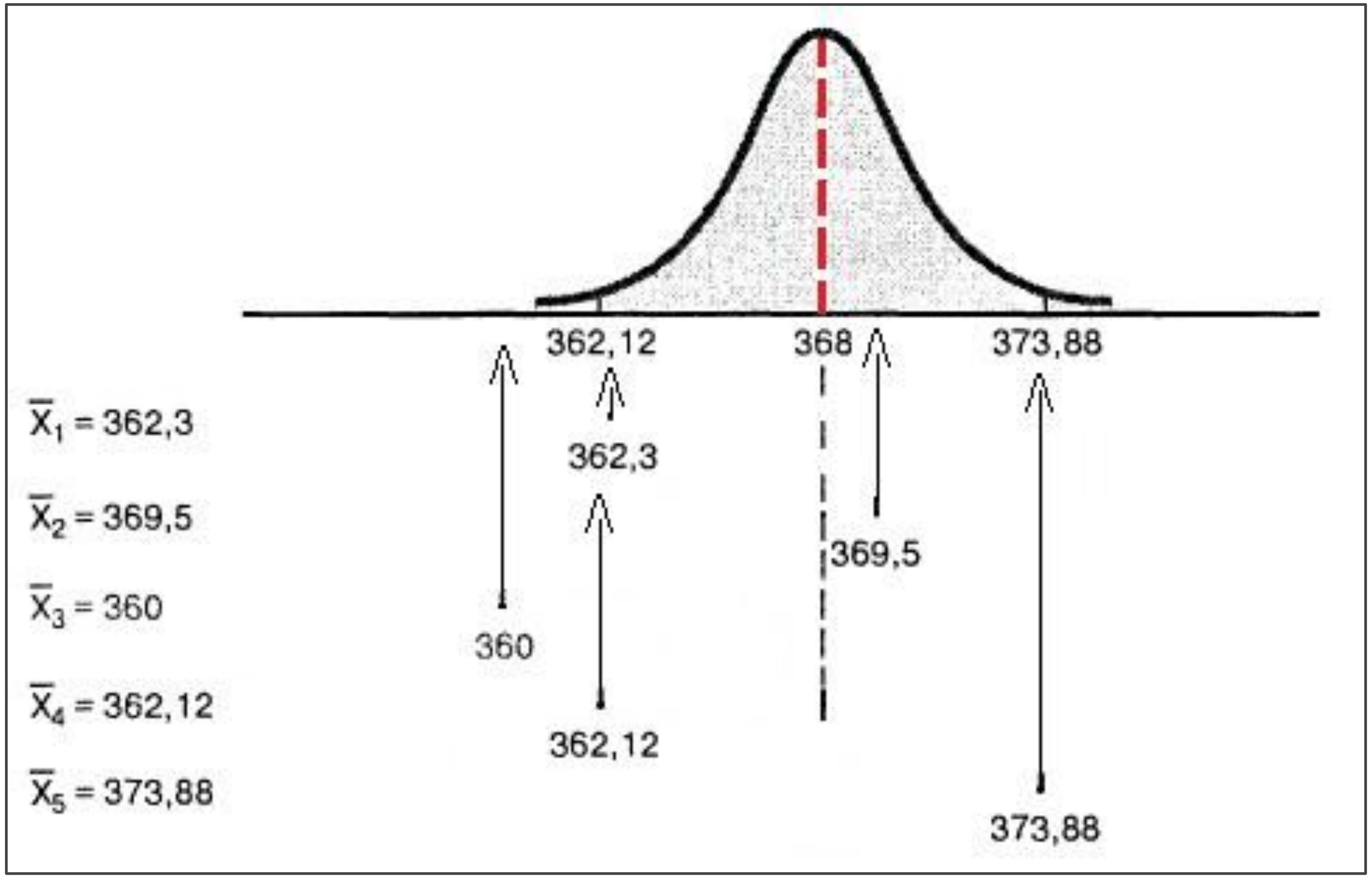
Пример: Мы провели опрос/голосование с целью выявить потенциального кандидата в президенты какой-либо страны. Опрос 1000 человек показал, что за этого кандидата проголосовало 362 человек - это 36,2%. Это лишь показатель по малой выборке по сравнению со всей страной. Поэтому такая оценка может не все говорить о реальной картине. Что в таком случае делать?!



Точечная оценка

Точечной оценкой называется число, которое используют для оценки параметра ГС (среднего).

Но ТО не так информативно, как ДИ!!!!



Что такое доверительный интервал?

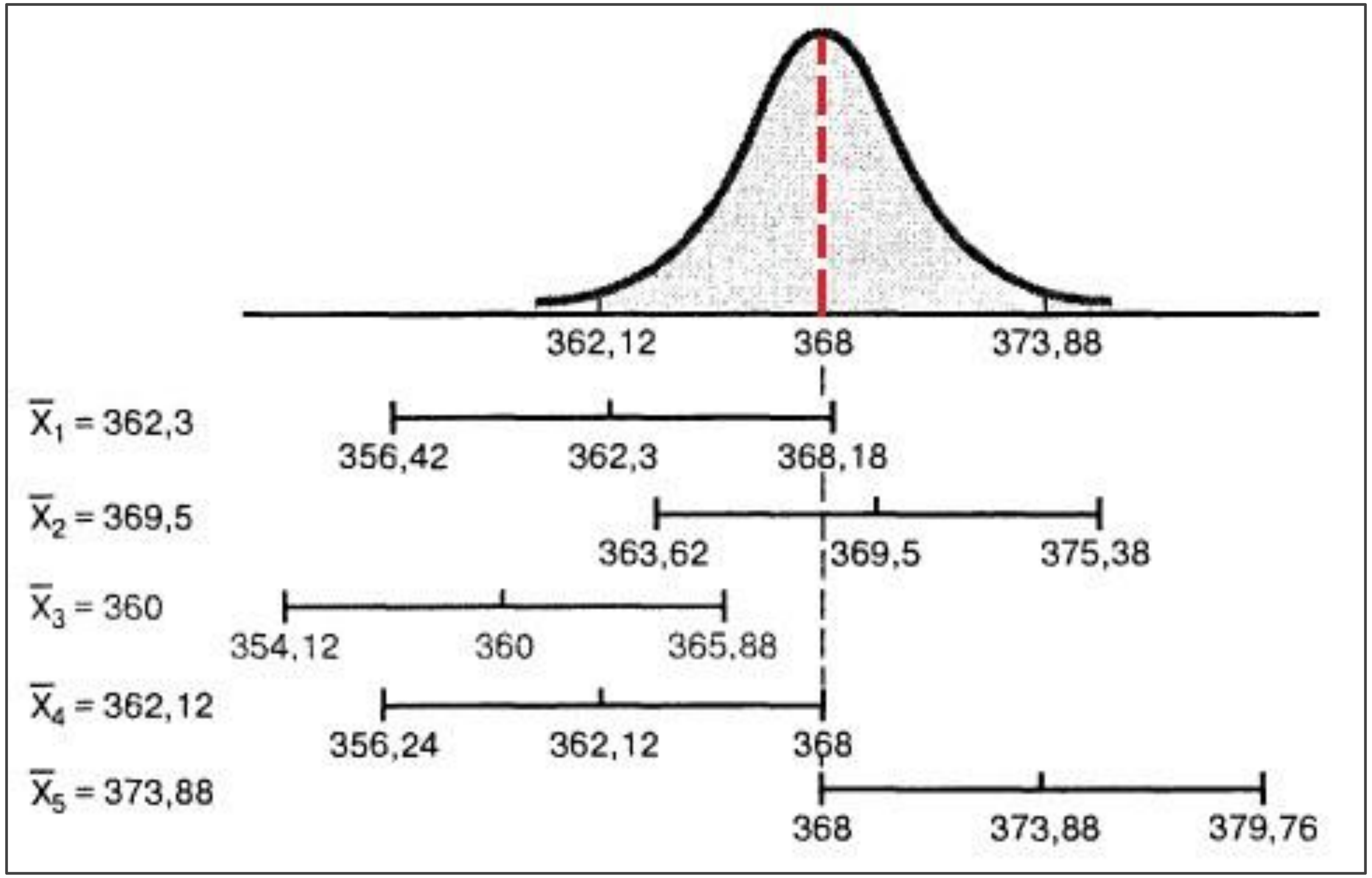
Цель: научиться оценивать параметры не просто числом, а целым интервалом, который покрывает возможные значения параметра с заданной вероятностью

Было бы намного лучше знать не просто оценку 36.2% , а интервал, в котором с большей вероятностью будет находится реальный процент сторонников нашего президента.
Скажем, с вероятностью 95% доля сторонников президента лежит в интервале от 35.6% до 36.8%

Точечная оценка VS доверительный интервал

Доверительный интервал - это интервал, который с заданной вероятностью накрывает оцениваемый нами параметр ГС.

Заметим, что ДИ для разных выборок одной и той же ГС могут отличаться!!!!

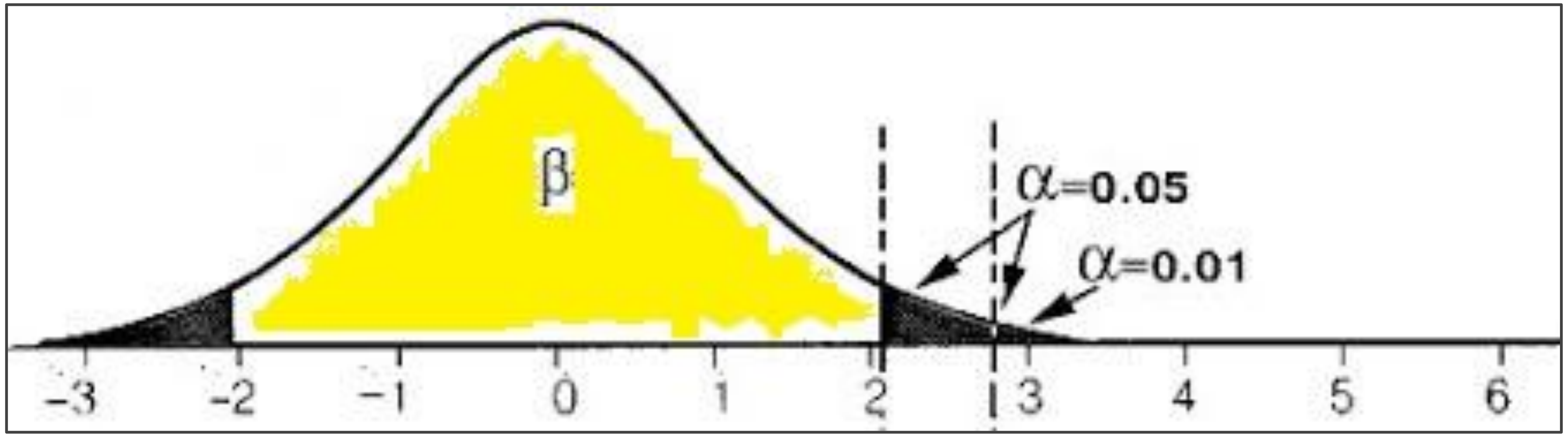


Уровень значимости и уровень доверия

→

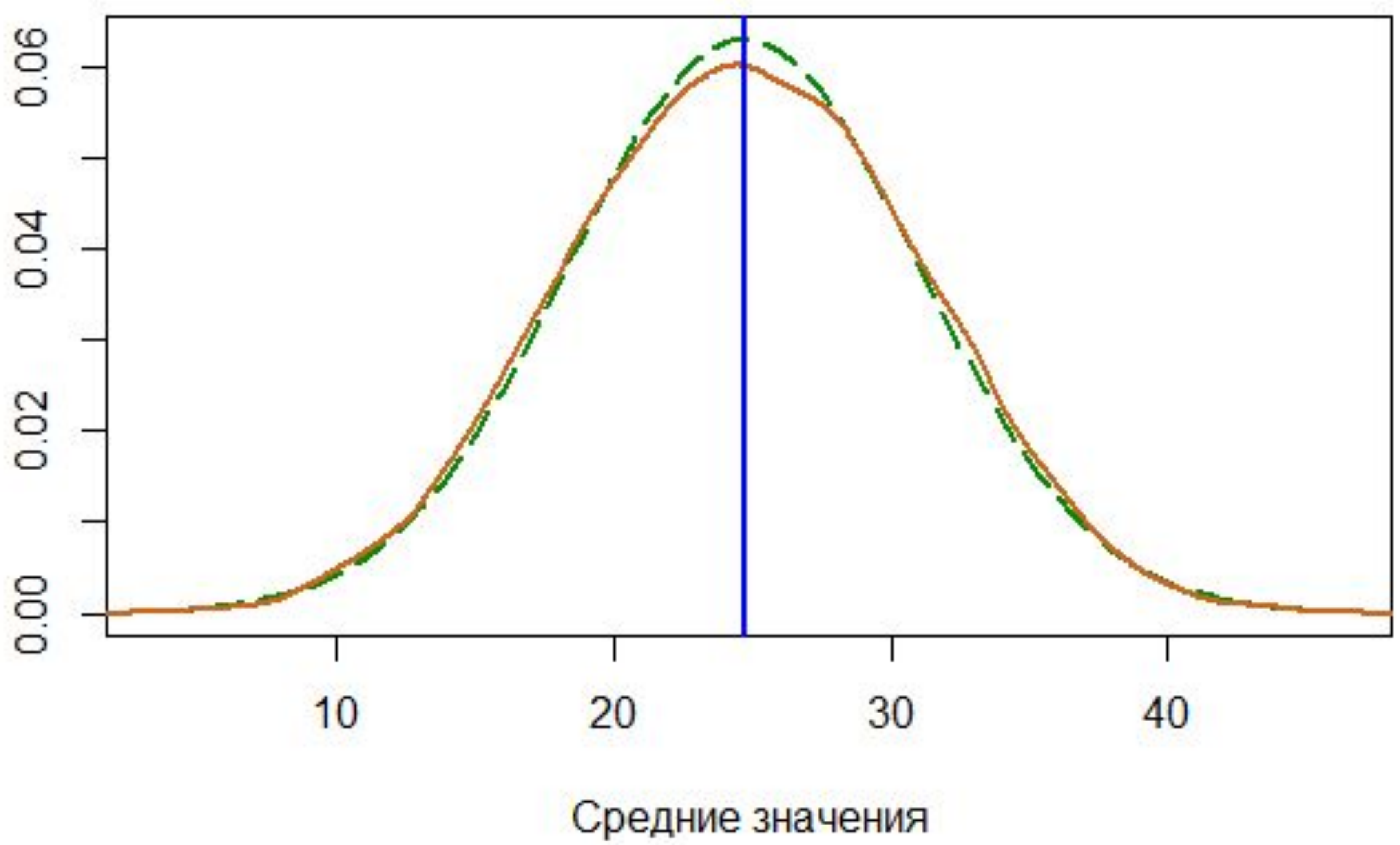
Уровень значимости α - это вероятность, с которой значение параметра не попадает в доверительный интервал.
Уровень доверия β = 1 − α - это вероятность того, что доверительный интервал накрывает значение параметра

Уровень значимости 0.01 и 0.05 соответствуют уровням доверия 0.99 и 0.95 соответственно. Величины могут выражаться в процентах т.е есть уровень доверия 0.99 и 99% - это одно и то же.



Условия для построения доверительного интервала

Теорема: Если распределение генеральной совокупности имеет конечные математическое ожидание и дисперсию, то при $n \to \infty$ основные выборочные характеристики (среднее, дисперсия, эмпирическая функция распределения) являются нормальными. Важно! Далее мы часто будем предполагать, что генеральная совокупность имеет нормальный закон распределения.



Подсчет доверительного интервала

Рассмотрим случайную выборку объема n, вычислим среднее значение $\overline{x}$ по выборке и зададим уровень значимости β.

Доверительный интервал для среднего имеет вид $(\overline{x} - \triangle; \overline{x} + \triangle)$,
где Δ это точность интервальной оценки.

Подсчет доверительного интервала

Вычисление **Δ** зависим о наших знание о ГС и с какой выборкой мы имеем дело.
Допустим нам известно стандартное отклонение σ генеральной совокупности.
Тогда $\triangle = \frac{a}{\sqrt{n}} z_a$, где Z_a это квантиль нормального распределения уровня 1-α/2.

Теорема:
Доверительный интервал для среднего с известной дисперсией имеет вид:

$$(\overline{x} - \frac{\sigma}{\sqrt{n}} z_a ; \overline{x} + \frac{\sigma}{\sqrt{n}} z_a)$$

Задача

Пример:
Дана выборка 9, 5, 7, 7, 4, 10, дисперсия σ^2 = 1.
Постройте 99% доверительный интервал.

$$(\bar{x} - \Delta; \bar{x} + \Delta)$$

$$\Delta = \frac{a}{\sqrt{n}} z_a$$

Задача

Задача:
Дана выборка 9, 5, 7, 7, 4, 10, дисперсия $\sigma^2 = 1$.
Постройте 99% доверительный интервал.

Решение:
Среднее значение равно $\bar{x} = (9+5+7+7+4+10) / 6 = 7$.
Доверительный интервал имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$.
По таблице нормального распределения находим
$1 - \alpha/2 = 0.995$ и определяем квантиль $z_\alpha = 2.58$.

Теперь можем найти точность $\Delta = (\sigma/\sqrt{n}) \cdot z_\alpha = 1/\sqrt{6} \cdot 2.58 \approx 1.05$
(здесь мы воспользовались тем, что известна дисперсия
генеральной совокупности). Искомый 99%-доверительный
интервал имеет вид $(7 - 1.05; 7 + 1.05) = (5.95; 8.05)$.

Задача

Задача:
(Похожие задачи буду в домашней работе)

Пример:
Пусть для выборки объема $n = 25$ вычислено среднее $\bar{x} = 130$. Из предыдущих исследований известно стандартное отклонение $\sigma = 12$. Постройте 98% доверительный интервал для среднего значения.

Решение:
Доверительный интервал имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$.
Уровень доверия равен $\beta = 0.98$, поэтому $\alpha = 0.02$.
По таблице нормального распределения находим
$1 - \alpha /2 = 0.99$ и определяем квантиль $z_\alpha = 2.33$.
Теперь можем найти точность $\Delta = (\sigma /\sqrt{n})^* z_\alpha = (12/\sqrt{25})^*2.33 \approx 5.59$. Искомый 98%-доверительный интервал имеет вид $(130 - 5.59; 130 + 5.59) = (124.41; 135.59)$.

Доверительный интервал для среднего при неизвестной дисперсии.

Что делать, если дисперсия нам по каким-то причинам неизвестна? Тогда мы можем посчитать ее вручную!

Важно, чтобы выборка была больше 30, тогда вместо σ мы будем использовать выборочное стандартное отклонение

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Теорема

Доверительный интервал для среднего при неизвестной дисперсии, но большой выборке ($n > 30$), имеет вид **$\bar{x} - (s/\sqrt{n})*z_\alpha; \bar{x} + s/\sqrt{n}*z_\alpha$** .

ДИ для среднего при неизвестной дисперсии и маленькой выборке.

Бывают случаи, когда выборка маленькая и про её параметры еще и мало чего известно. Тогда в таком случае вместо нормального распределения используем t-распределение.

Формула: но в формуле процентиль z_α заменяем на t_α(n-1) - это квантиль распределения Стьюдента уровня 1-α/2 со степенью свободы = n-1. (это число дано в таблице распределения)

Степень свободы

→ Степенью свободы = n-1

Значение t-критерия Стьюдента при уровне значимости
(0,01; 0,05; 0,1; 0,15; 0,20; 0,25; 0,30)

Таблица 1

n	P - 0,01	P - 0,05	P - 0,1	P - 0,15	P - 0,2	P - 0,25	P - 0,3
1	63,6567412	12,7062047	6,3137515	4,1652998	3,0776835	2,4142136	1,9626105
2	9,9248432	4,3026527	2,9199856	2,2819306	1,8856181	1,6035675	1,3862066
3	5,8409093	3,1824463	2,3533634	1,9243197	1,6377444	1,4226253	1,2497781
4	4,6040949	2,7764451	2,1318468	1,7781922	1,5332063	1,3443976	1,1895669
5	4,0321430	2,5705818	2,0150484	1,6993626	1,4758840	1,3009490	1,1557673
6	3,7074280	2,4469119	1,9431803	1,6501732	1,4397557	1,2733493	1,1341569
7	3,4994833	2,3646243	1,8945786	1,6165917	1,4149239	1,2542787	1,1191591
8	3,3553873	2,3060041	1,8595480	1,5922214	1,3968153	1,2403183	1,1081454
9	3,2498355	2,2621572	1,8331129	1,5737358	1,3830287	1,2296592	1,0997162
10	3,1692727	2,2281389	1,8124611	1,5592359	1,3721836	1,2212554	1,0930581

Что за степени свободы и кто такой Стьюдент?

Замечание: Число степеней свободы зависит от того, сколько имеется связей между наблюдениями. Так как мы знаем среднее, то наблюдения связаны одним равенством и степеней свободы становится на одну меньше. То, что других связей нет, надо доказывать, но их действительно нет. Честное слово.

Замечание: Распределение Стьюдента было введено в 1908 году В.С. Госсетом, ирландским служащим пивоваренного завода, который участвовал в разработке новых технологий производства пива и никаким студентом не был. Придавать известности результаты исследований означало открыть корпоративную тайну, поэтому Госсет напечатал свои материалы под псевдонимом Стьюдент. Фишер ввел для него обозначение t-распределение

Пример для подсчета ДИ в случае маленькой выборки:

Задача: Пусть объем выборки $n = 16$, выборочное среднее $\bar{x} = 5$, выборочная дисперсия $s^2 = 4$. Постройте 99% доверительный интервал.

Решение: Среднее значение равно $\bar{x} = 5$, а выборочная дисперсия $s^2 = 4$.
Так как неизвестна дисперсия генеральной совокупности и $n < 30$, поэтому точность интервальной оценки $\Delta = s/\sqrt{n} \cdot t_\alpha$. По таблице распределения Стьюдента находим $1 - \alpha/2 = 0.995$ и, так как у нас $n - 1 = 16 - 1 = 15$ степеней свободы, определяем квантиль $t_\alpha = 3.29$. Теперь можем найти точность $\Delta = (s/\sqrt{n}) \cdot t_\alpha = (2/\sqrt{16}) \cdot 3.29 \approxeq 1.645$.
Искомый 99%-доверительный интервал имеет вид $(5 - 1.645;\ 5 + 1.645) = (3.355;\ 6.645)$.

Вернемся к нашим президентским выборам:

→

Скольких людей надо опросить, чтобы наши результаты давали 95% точность?

Обратная задача: раз мы знаем ДИ, ток можно ли найти минимальный объем выборки голосовавших, для того, чтобы с заданной точностью и уровнем доверия найти среднее.

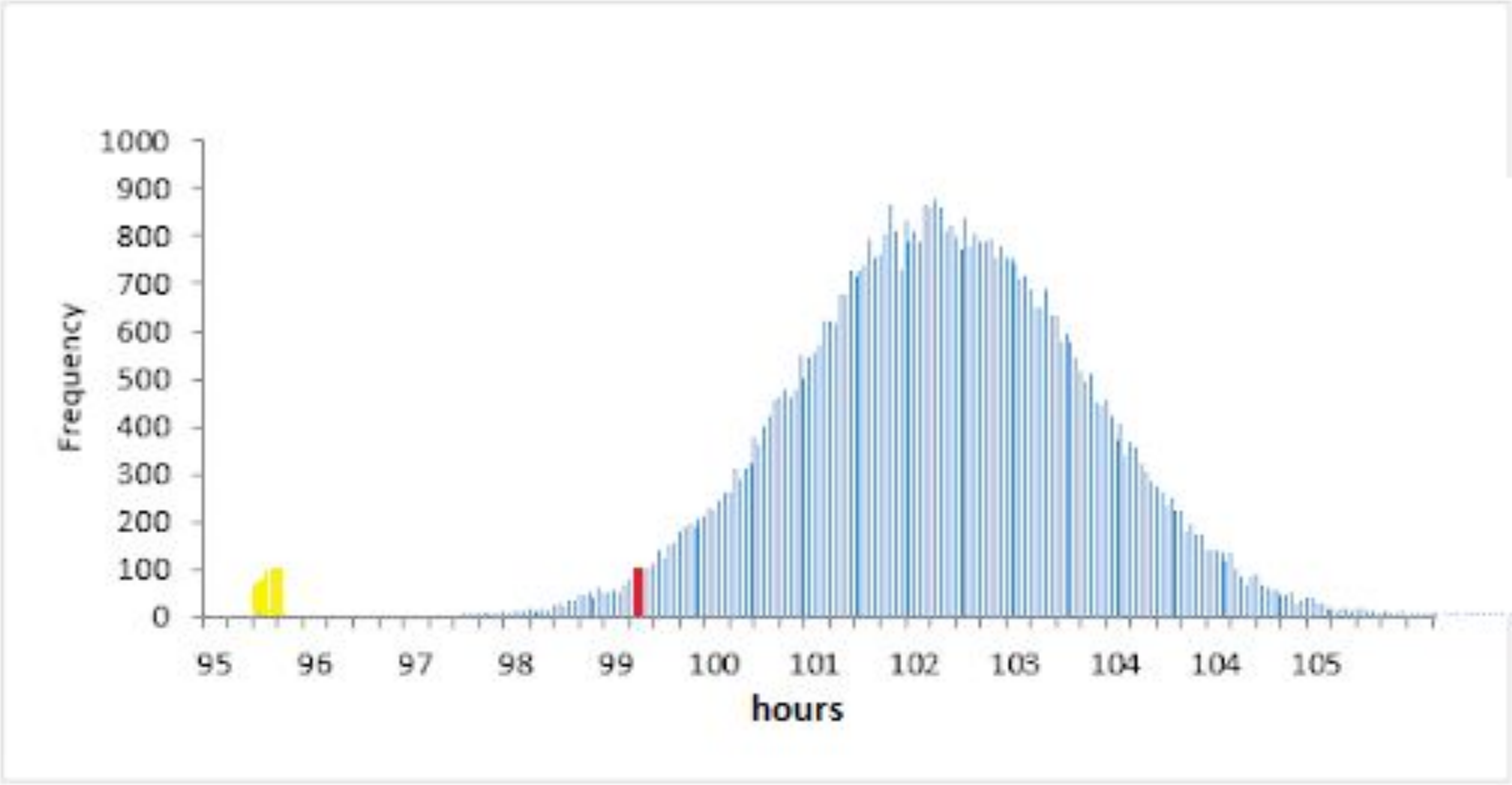
Важно! Для того чтобы найти минимальный необходимый объем выборки для построения доверительного интервала для среднего значения с заданной точностью Δ и уровнем значимости α, достаточно применить формулу.

$$n = \left(\frac{z_a \sigma}{\Delta}\right)^2$$

Теперь понятно, как определить объем выборки при проведении собственных исследований!

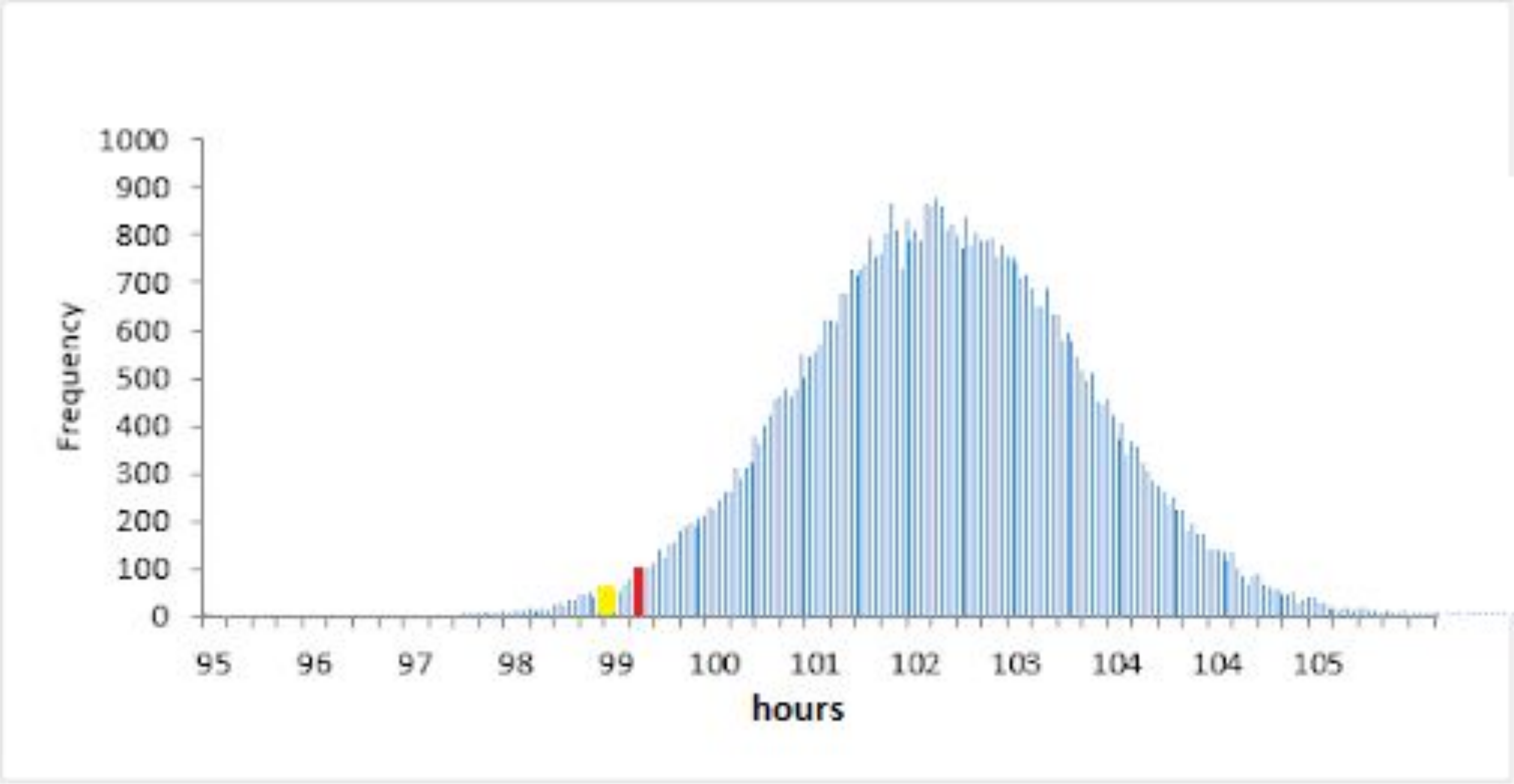
Проверка гипотез

Что можно сказать о результативности препарата?



Проверка гипотез

Что можно сказать о результативности препарата теперь?



Проверка гипотез

Этот процесс известен как **проверка гипотез** (проверка статистических гипотез или проверка значимости).

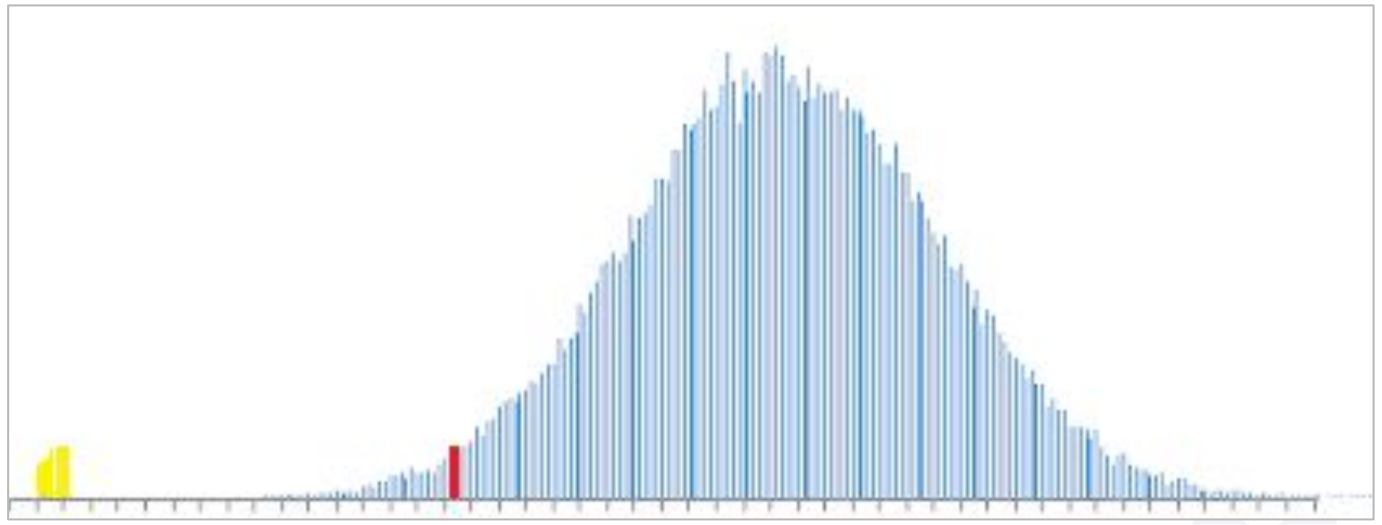
Замечание: Нулевая гипотеза всегда относится к популяции, представляющей больший интерес, нежели выборка. В рамках проверки гипотезы мы либо отвергаем нулевую гипотезу и принимаем альтернативу, либо не отвергаем нулевую гипотезу.

Логика проверки гипотез

Всегда составляют и проверяют нулевую гипотезу **(H_0)**, которая отвергает эффект (например, эффект случайный) в генеральной совокупности. Например, при времени выздоровления пациентов - нулевая гипотеза **H_0** означала бы, что показатели времени выздоровления не отличаются от времени выздоровления без лекарства.

Затем определяют и проверяют альтернативную гипотезу **(H_1)**, которая принимается, если нулевая гипотеза неверна. Альтернативная гипотеза больше относится к той теории, которую собираются исследовать. На примере альтернативная гипотеза **(H_1)** заключается в утверждении, что препарат помогает ускорить выздоровление.

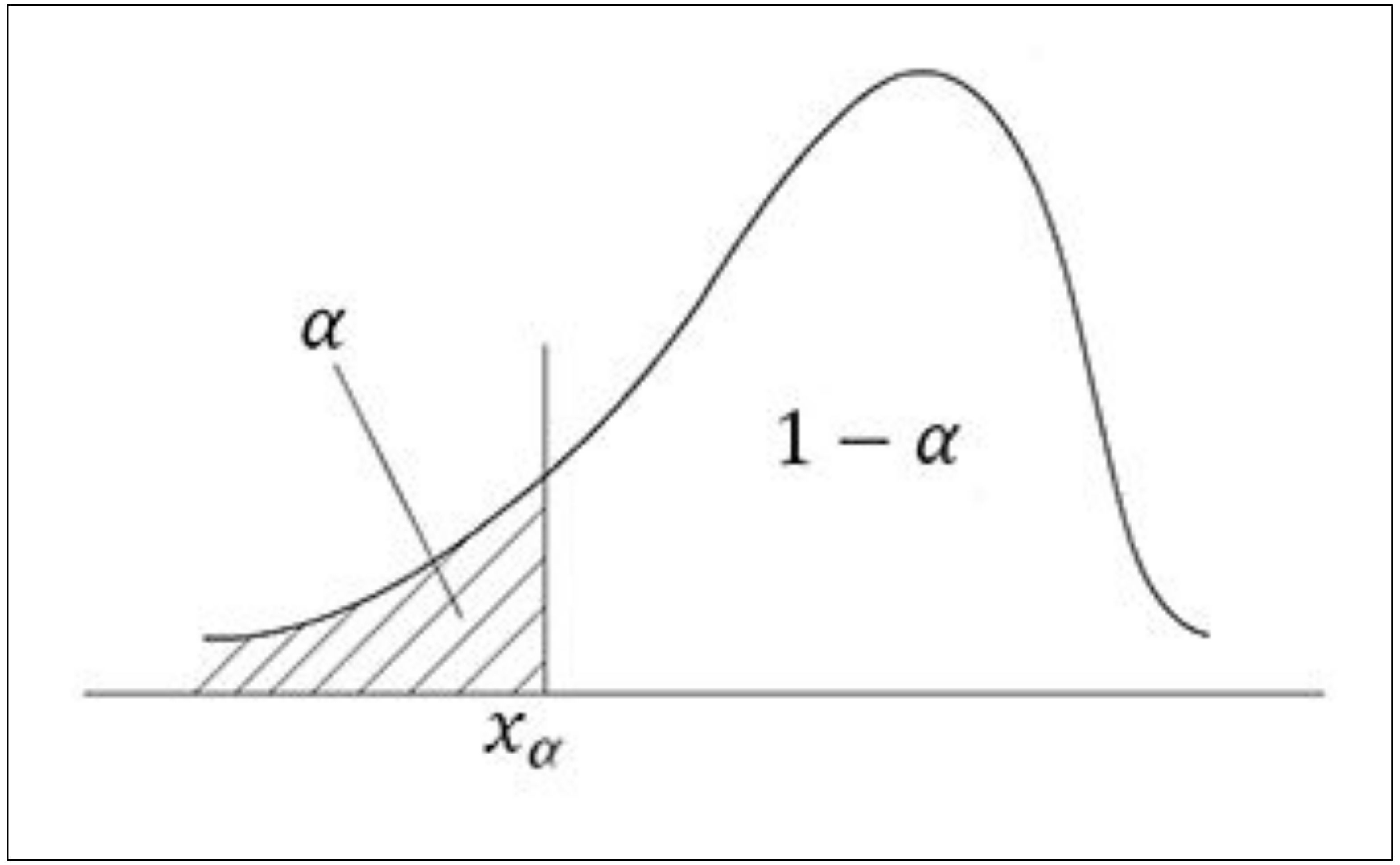
Уровень значимости. Важным этапом проверки статистических гипотез является определение уровня *статистической значимости* , т.е. максимально допускаемой исследователем вероятности ошибочного отклонения нулевой гипотезы.



Проверка гипотез

α - заданный исследователем уровень значимости.

Если наблюдаемое значение критерия (***K***) принадлежит критической области ($K_{кр}$), заштрихованная область на рисунке), гипотезу отвергают, если не принадлежит - не отвергают.



Для краткости можно записать и так:

$|K| > K_{кр}$ - отклоняем H_0

$|K| < K_{кр}$ - не отклоняем H_0

После того как данные будут собраны, значения из выборки подставляют в формулу для вычисления статистики критерия. (об этом чуть позже)

Эта величина количественно отражает аргументы в наборе данных против нулевой гипотезы.

Применение значения P-value

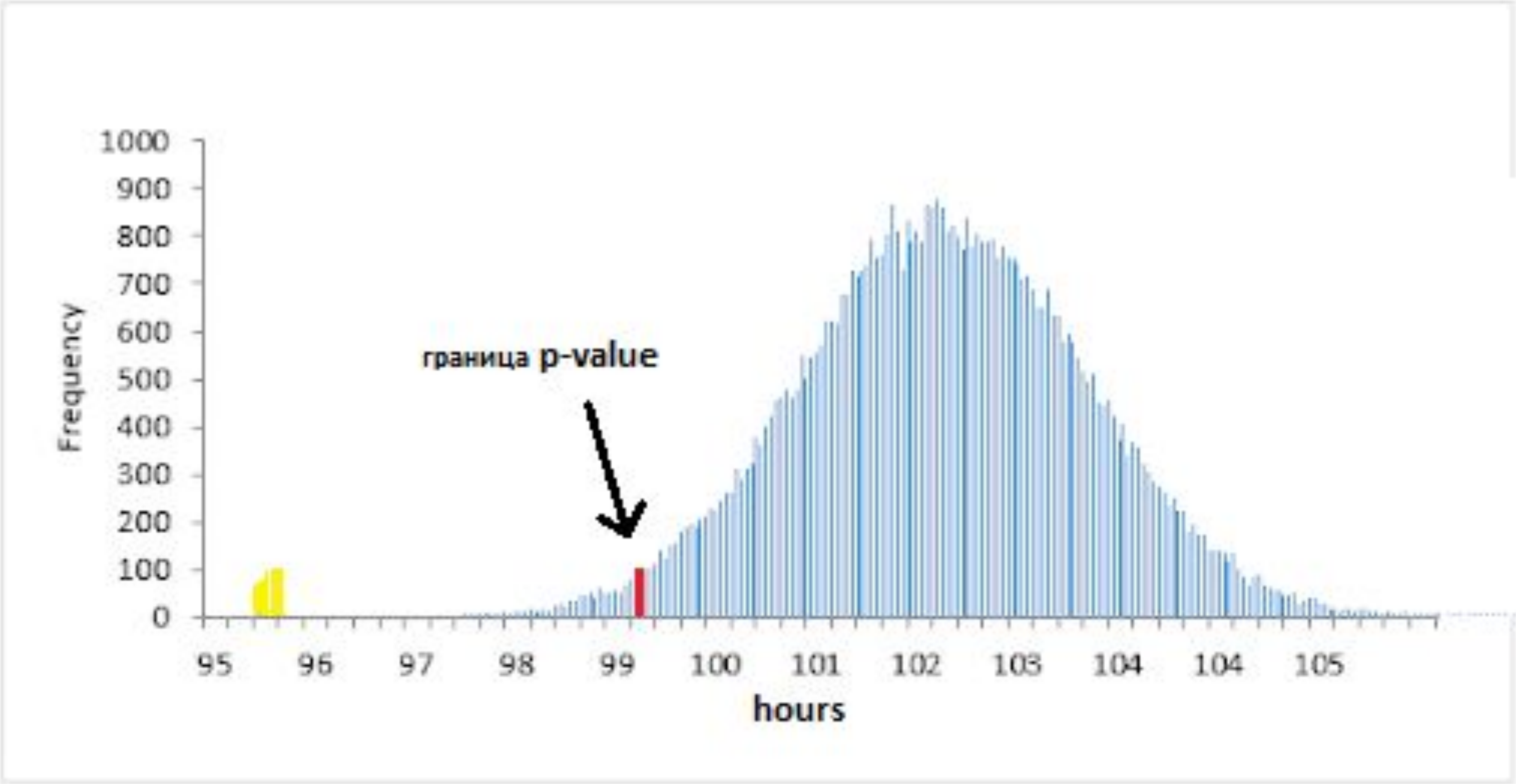
Значение статистики критерия, полученное из выборки, связывают с уже известным распределением, которому она подчиняется, чтобы получить значение ***p***, площадь обоих "хвостов" (или одного "хвоста", в случае односторонней гипотезы) распределения вероятности.

Большинство компьютерных пакетов обеспечивают автоматическое вычисление двустороннего значения ***p***. (Python)

Определение:

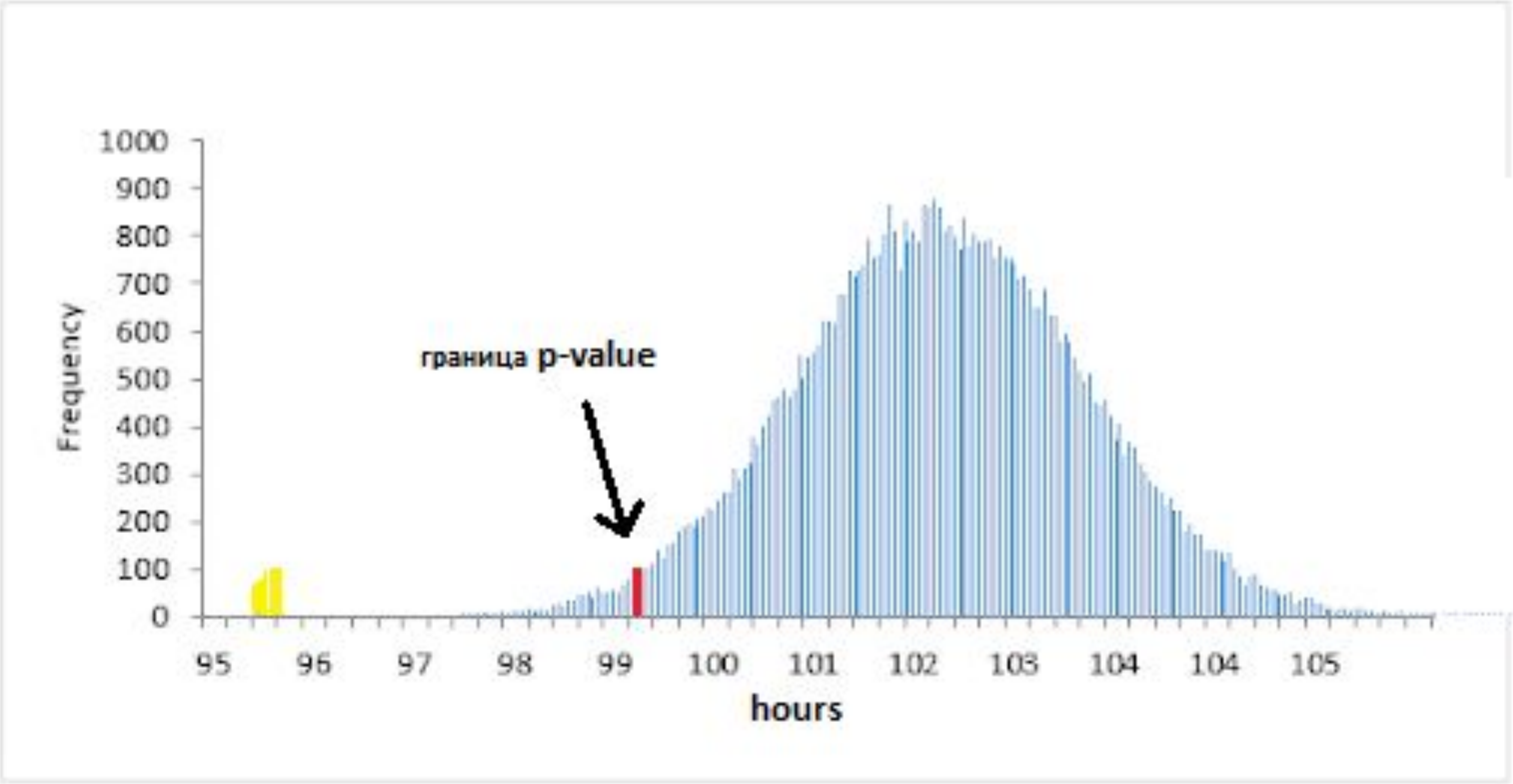
Значение ***p*** — это вероятность получения нашего вычисленного значения или его еще большего значения, если нулевая гипотеза верна.

Иными словами, ***p*** - это вероятность отвергнуть нулевую гипотезу при условии, что она верна.



Применение значения P-value

Можно ли по графику сказать, что наше лекарство помогает в лечении?



Низкие значения p - это хорошо; Они указывают на то, что ваши данные не были случайными.

Логика применения значения P-value

Следует решить, сколько аргументов позволяют отвергнуть нулевую гипотезу в пользу альтернативной. Чем меньше значение *p*, тем сильнее аргументы против нулевой гипотезы.

Традиционно полагают, если $p < 0,05$, $(=0,05)$ то аргументов достаточно, чтобы отвергнуть нулевую гипотезу, хотя есть небольшой шанс против этого. Тогда можно отвергнуть нулевую гипотезу и сказать, что результаты значимы на 5% уровне. Если это может привести к серьезным последствиям, необходимо потребовать более веских аргументов, прежде чем отвергнуть нулевую гипотезу, например, выбрать значение = 0,01 (или 0,001).

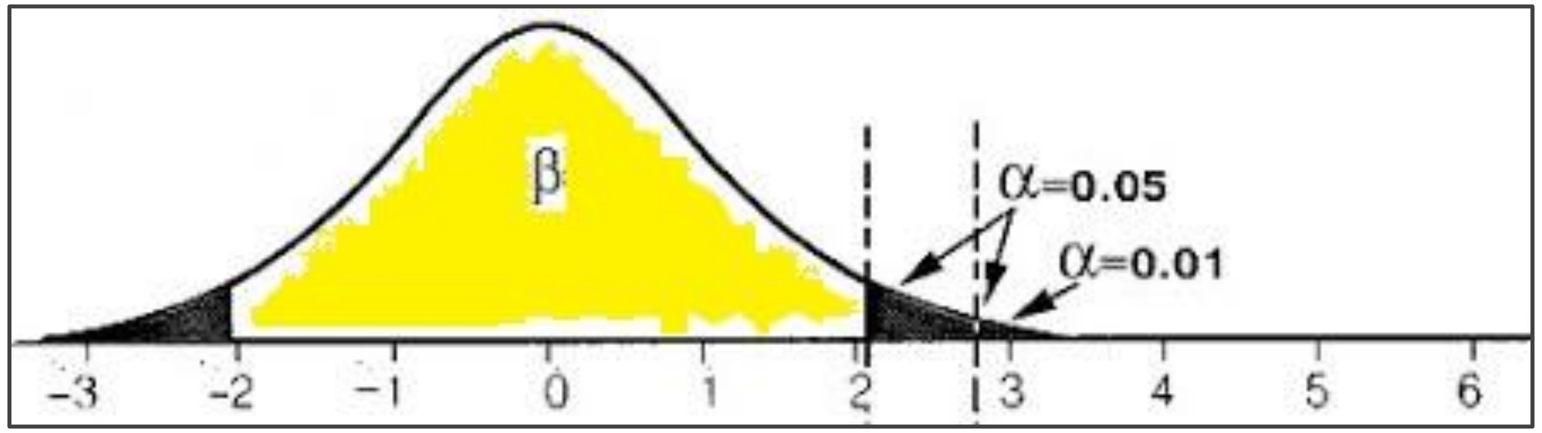
Напротив, если $p > 0,05$, то аргументов недостаточно, чтобы отвергнуть нулевую гипотезу. Не отвергая нулевую гипотезу, можно заявить, что результаты не значимы на 5% уровне. Данное заключение не означает, что нулевая гипотеза истинна, просто недостаточно аргументов (возможно, маленький объем выборки), чтобы ее отвергнуть.

Проверка гипотез против доверительных интервалов

Доверительные интервалы и проверка гипотез тесно связаны. Первоначальная цель проверки гипотезы состоит в том, чтобы принять решение и предоставить точное значение **p**.

ДИ предоставляют интервал вероятных значений для истинного эффекта, поэтому его также можно использовать для принятия решения даже без точных значений **p**.

Например, если бы гипотетическое значение для данного эффекта (например, значение, равное нулю) находилось вне 95% ДИ, можно было бы счесть гипотетическое значение неправдоподобным и отвергнуть. В этом случае станет известно, что **p** < 0,05, но не станет известно его точное значение.



Проверим себя

В нашем исследовании **p** уровень значимости равен 0.003

- Вероятность того, что верна нулевая гипотеза (новый препарат не влияет на скорость выздоровления) также равняется всего-лишь 0,003.
- Если бы в исследовании мы получили **p**=0.9, это означало бы, что верна нулевая гипотеза, и новый препарат не влияет на скорость выздоровления.
- Чем меньше **p** уровень значимости, тем сильнее получаемые различия. Например, если бы **p** уровень значимости в нашем исследовании был бы равен 0.0001, значит новый препарат еще сильнее бы влиял на скорость выздоровления.
- Статистически значимый результат, всегда означает ценный и осмысленный результат.
- Вероятность случайно получить такие различия равняется 0.003.

Проверим себя

Использование доверительных интервалов зачастую рассматривают, как альтернативный способ проверки гипотез. В нашем случае, если значение 20 (предполагаемое среднее значение в генеральной совокупности) не будет принадлежать 95% доверительному интервалу, рассчитанному по выборочным данным, у нас будет достаточно оснований отклонить нулевую гипотезу.

Проверьте согласуются ли результаты двух этих подходов: рассчитайте 95% доверительный интервал для среднего значения, на примере с тестированием нового препарата.

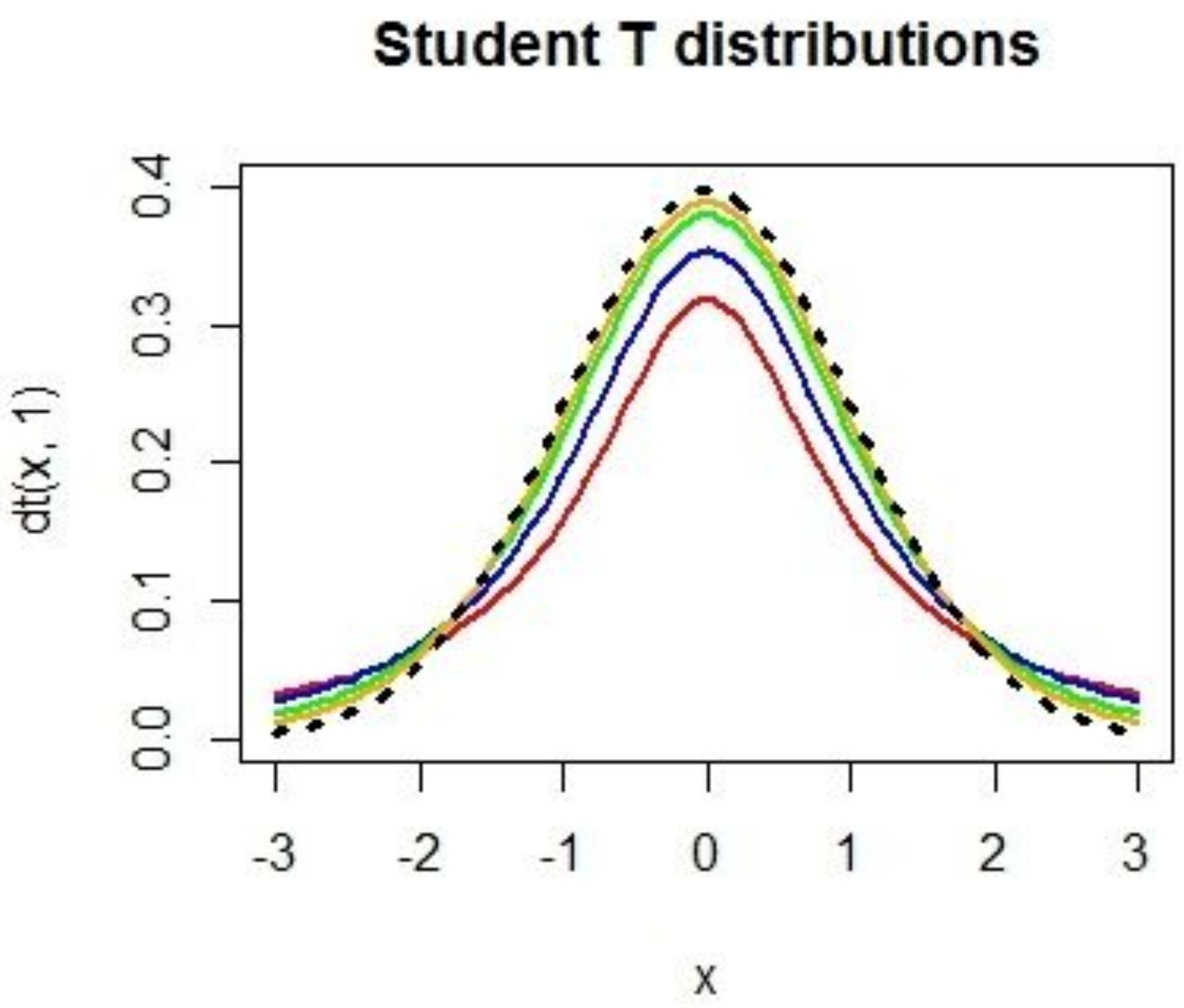
n = 64, sd = 4, M = 18.5

Выбери верное утверждение:

- 20 принадлежит доверительному интервалу — отклоняем H1
- 20 не принадлежит доверительному интервалу — отклоняем H1
- 20 принадлежит доверительному интервалу — отклоняем H0
- 20 не принадлежит доверительному интервалу — отклоняем H0

Тесты проверки гипотез

t- тест: С помощью t-теста (также называемого Т-тестом Стьюдента) сравниваются два средних значения (средних) и сообщается, отличаются ли они друг от друга. Т-тест также показывает, насколько значительны различия; Другими словами, он позволяет узнать, могли ли эти различия возникнуть случайно.



Тесты проверки гипотез

Если мы хотим сравнить распределение категориальных переменных, то обычно строят таблицы сопряженности и используют критерий χ^2.

Например, из интервью с носителями одной деревни произвольным образом выбрали по пол часа и посчитали кол-во реализаций диалектных форм vs. недиалектных. В результате получилось что у женщин было 107 диалектных форм vs. 93 недиалектные, а у мужчин — 74 vs. 45. Значима ли зафиксированная разница?

На самом деле тестов больше, но мы пока остановимся на самых важных.

Если одно из ожидаемых значений меньше 5, то следует использовать тест Фишера:

107	74
93	45

А как же Python?

Примеры на python
Примеры на python
Примеры на python
Примеры на python

Спасибо за внимание!

Миленькин Александр
Биоинформатик в Insilico Medicine

fb.com/

mail@mail.com

нетология