

Доверительные интервалы

Основы биостатистики, осень 2022

Марина Варфоломеева

Доверительные интервалы. Основы тестирования гипотез.

- Выборочное распределение и распределение выборочных средних
- Центральная предельная теорема (ЦПТ)
- Стандартная ошибка среднего
- Стандартизация
- Стандартное нормальное распределение
- Доверительный интервал для среднего из нормального и t -распределения
- Как устроено тестирование гипотез
- Одновыборочный t -тест
- Ошибки при тестировании гипотез

Выборочная оценка среднего

Как можно судить о свойствах генеральной совокупности по выборке?

Центральная предельная теорема (ЦПТ) говорит, что если мы возьмем достаточно большую выборку из генеральной совокупности, то среднее значение будет нормально распределено с параметрами $\mu_{\bar{x}}$ и $\sigma_{\bar{x}}$:

$$\bar{X} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

При чем $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.

Но самое важное, что при больших объемах выборки ($N > 30$, или даже $N > 100$) это так, даже если x в генеральной совокупности не подчиняется нормальному распределению.

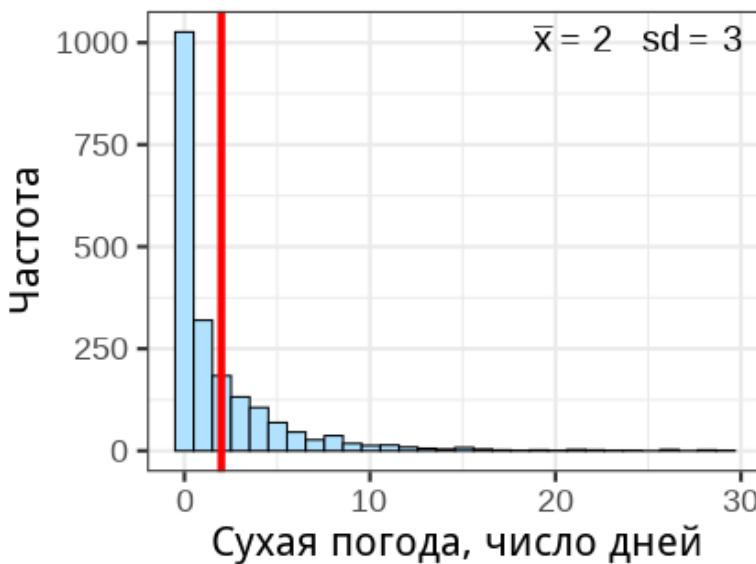
Продолжительность сухой погоды

В аэропорту Виннипега (Канада) в течение нескольких лет регистрировали продолжительность периодов погоды без осадков (в днях). Представим, что данные из датасета `droughts` (пакет `DAAG`) — это генеральная совокупность.

Перед вами распределение продолжительности периодов без осадков (число дней).

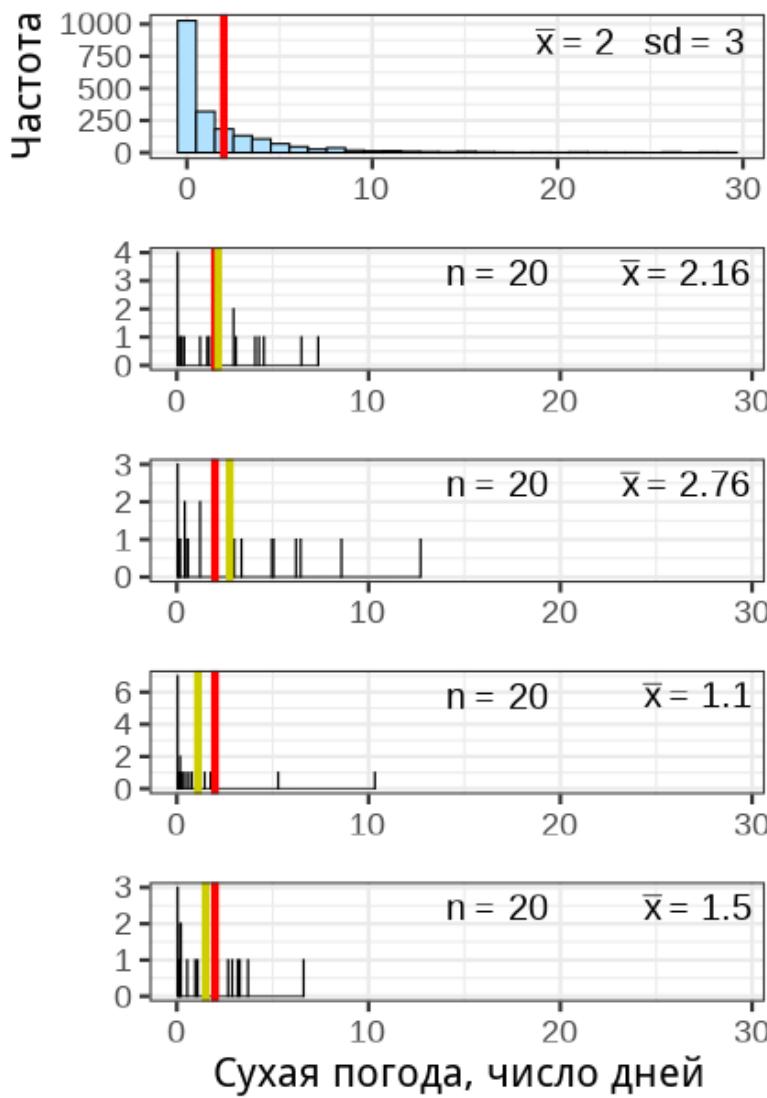
Мы хотим оценить среднюю продолжительность сухой погоды.

Давайте будем брать из нашей “генеральной совокупности” выборки и оценивать по ним среднее значение.



old winnipeg airport by twostoutmonks on Flickr

Средние в выборках

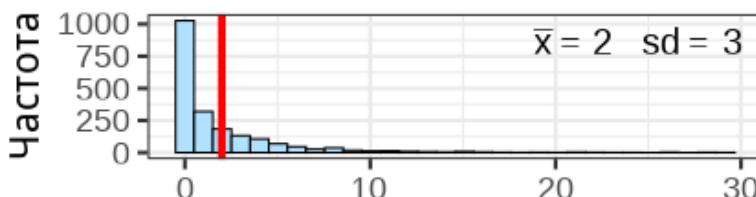


Средние в выборках отличаются от среднего в генеральной совокупности.

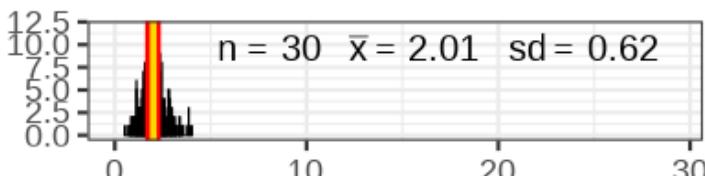
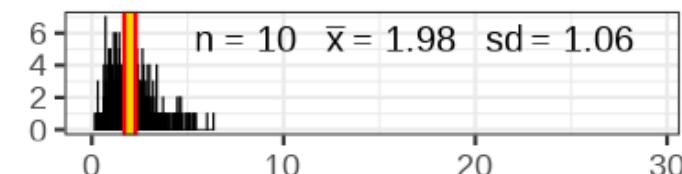
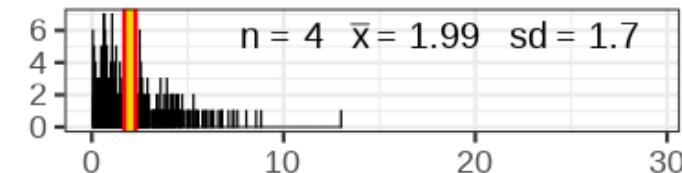
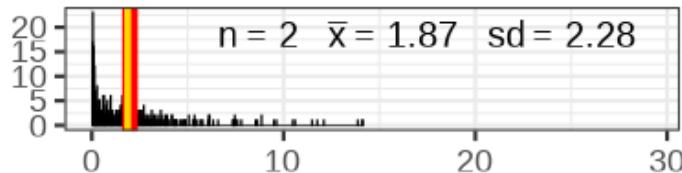
Если взять много выборок определенного размера, можно построить распределение выборочных средних.

Давайте посмотрим, как меняется форма распределения выборочных средних при изменении объема выборки.

Распределение выборочных средних



$$\bar{X} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$$



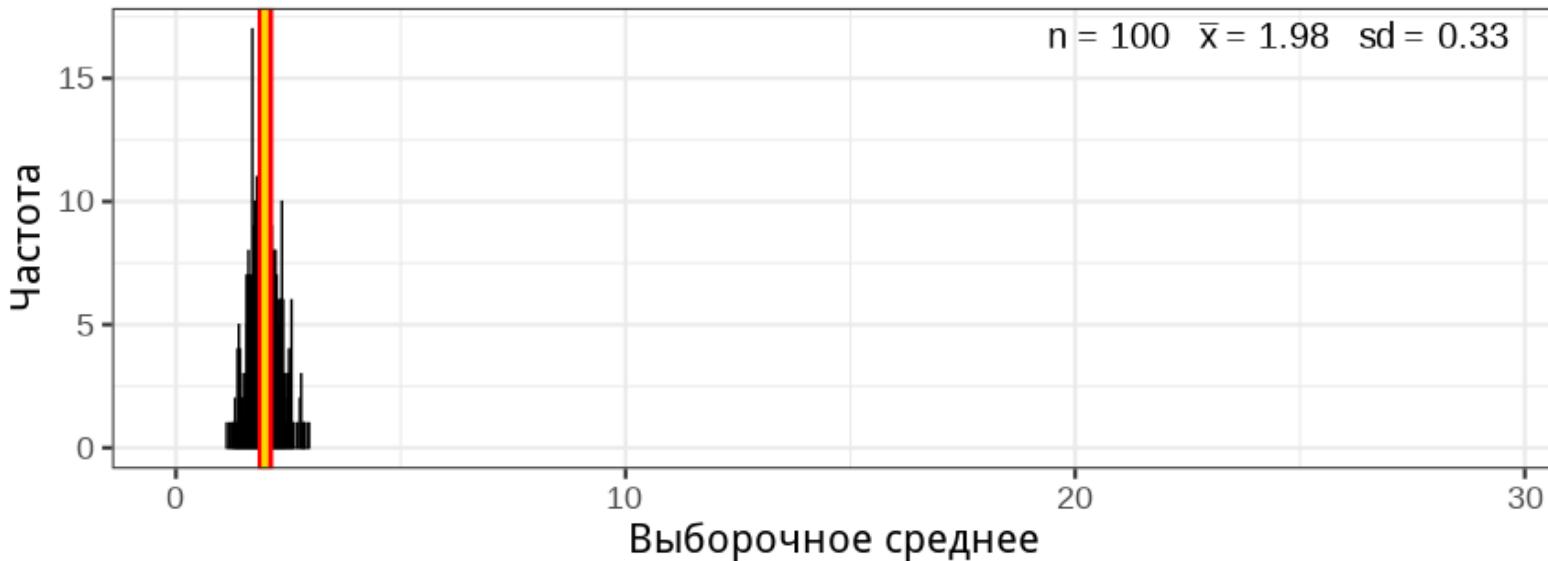
Выборочное среднее

$\mu_{\bar{x}} = \mu$ — среднее значение выборочных средних стремится к среднему в генеральной совокупности.

$\sigma_{\bar{x}} = \sigma / \sqrt{n}$ — стандартное отклонение в \sqrt{n} раз меньше стандартного отклонения в генеральной совокупности.

$\sigma_{\bar{x}}$ называется **стандартная ошибка среднего** и обозначается $SE_{\bar{x}}$.

Центральная предельная теорема очень важна в статистике



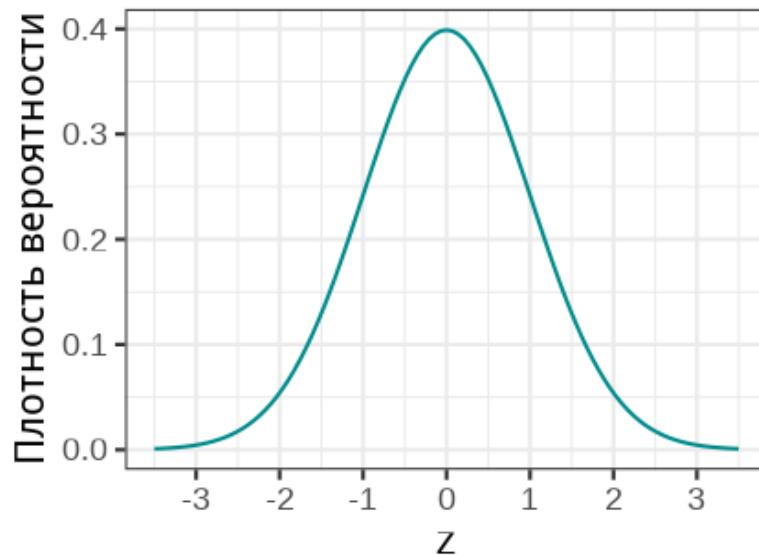
$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

Пользуясь ее выводами, мы сможем:

- строить доверительные интервалы
- тестировать гипотезы

Стандартизация

Стандартное нормальное распределение



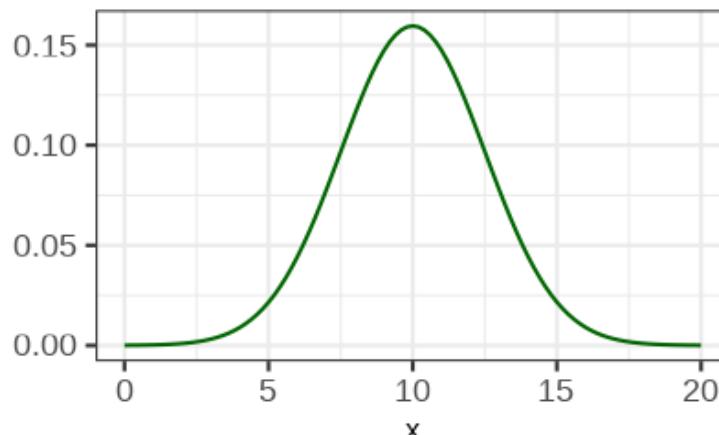
$$N(0, 1)$$

Нормальное распределение со средним значением $\mu = 0$ и стандартным отклонением $\sigma = 1$ называется **стандартное нормальное распределение**.

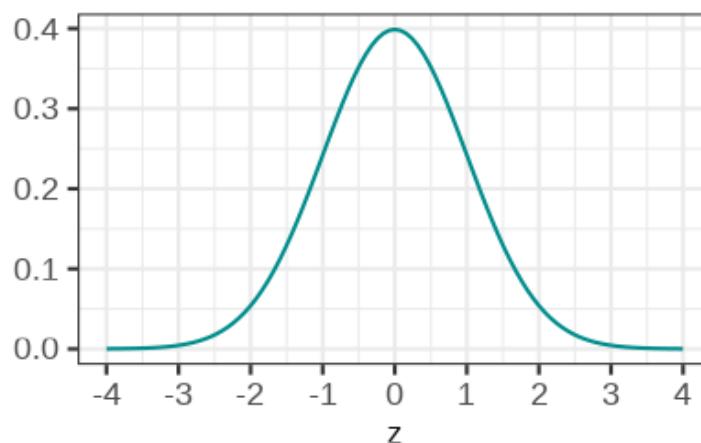
Ось X на графике такого распределения фактически измерена в стандартных отклонениях.

Стандартизация (Z-преобразование)

До стандартизации



После стандартизации



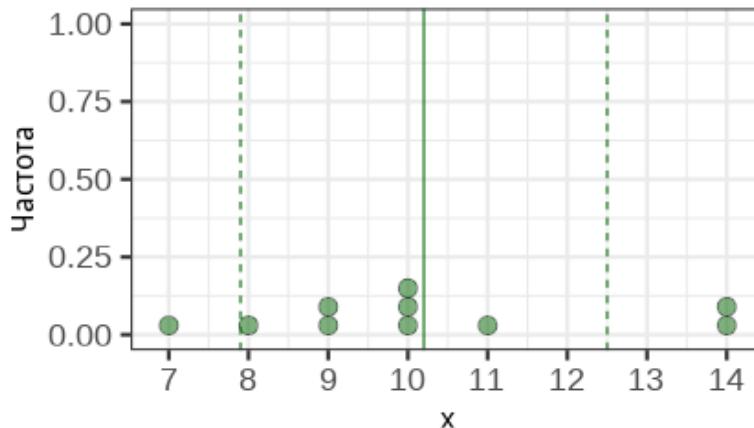
$$z = \frac{x - \mu}{\sigma}$$

После стандартизации любое нормальное распределение $X \sim N(\mu, \sigma)$ превращается в стандартное $Z \sim N(0, 1)$.

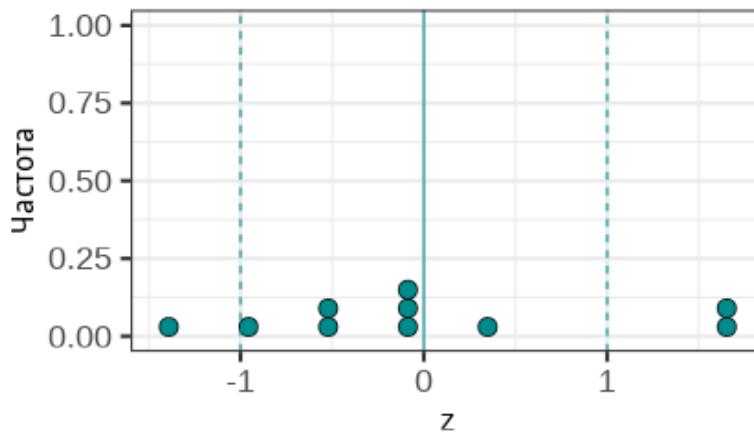
Для стандартного нормального распределения легко можно посчитать вероятность (площадь под кривой). Раньше для него составляли статистические таблицы.

Стандартизацию можно применять не только к нормально-распределенным величинам

До стандартизации $\bar{x} = 10$, $s = 2.30$



После стандартизации $\bar{z} = 0$, $s_z = 1$



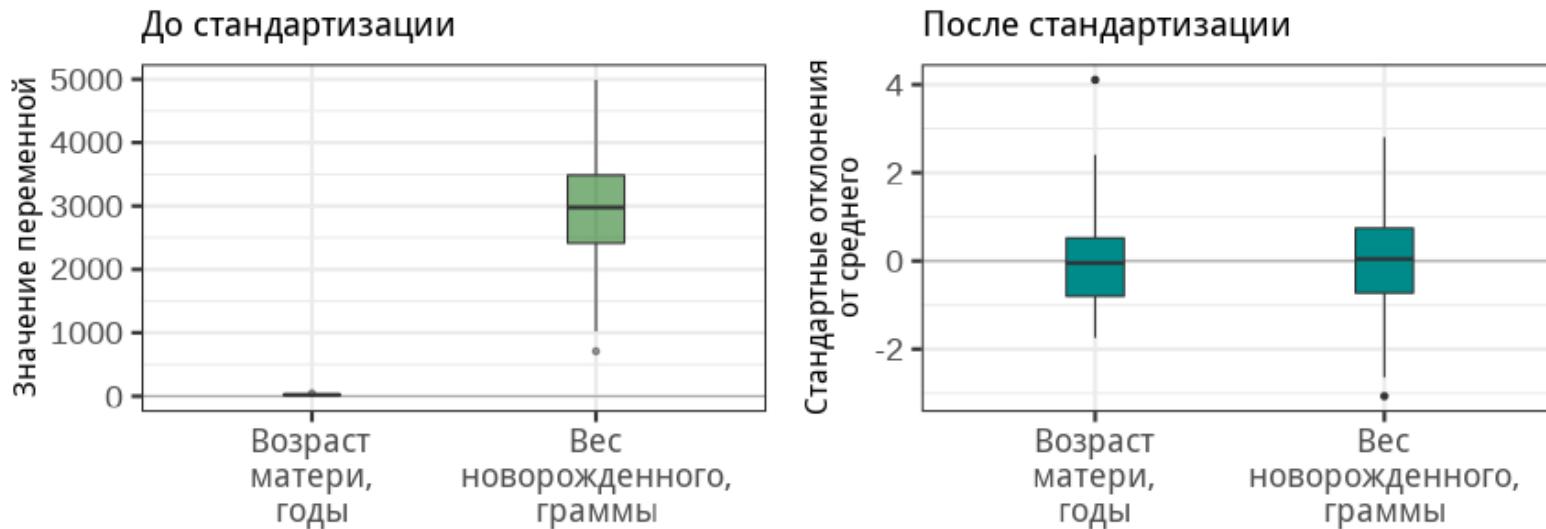
$$z_i = \frac{x_i - \bar{x}}{s}$$

Стандартизованная величина (Z-оценка) показывает, на сколько стандартных отклонений значение отличается от среднего

После стандартизации всегда:

- среднее $\bar{z} = 0$
- стандартное отклонение $s_z = 1$

Стандартизация позволяет уравнять шкалы, в которых измерены переменные

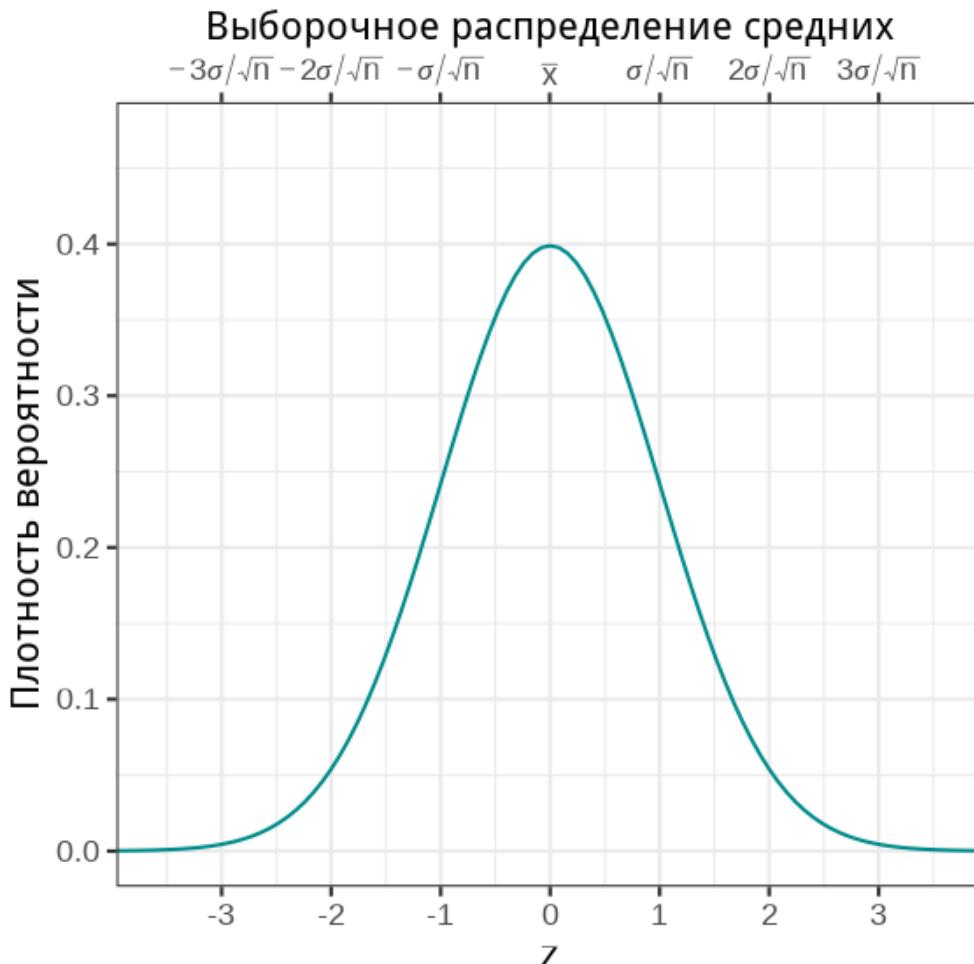


Стандартизованные данные используются

- для сопоставления величин, изначально измеренных в разных шкалах
- для уменьшения абсолютной величины значений в сложных вычислениях

Доверительный интервал

Если выполняется центральная предельная теорема, то



$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

После стандартизации:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Т.е. **стандартизованное распределение выборочных средних** будет подчиняться **стандартному нормальному распределению**.

Доверительный интервал



Доверительный интервал — это интервал, в который попадает заданный процент выборочных средних.

В 95% повторных выборок выборочное среднее попадет в 95% доверительный интервал.

Смысл 95% доверительного интервала



- Красная вертикальная линия — среднее в генеральной совокупности. Это фиксированная величина (она либо попала в интервал, либо нет).
- Горизонтальные линии — доверительные интервалы в 100 случайных выборок из одной и той же генеральной совокупности.

В $\approx 95\%$ повторных выборках (того же размера) доверительные интервалы “накроют” истинное среднее значение.

Способы расчета доверительного интервала

Доверительный интервал из нормального распределения



Условия применимости нормального распределения для доверительного интервала

1. Должна быть известна σ в генеральной совокупности.
2. Должны выполняться условия, при которых справедлива ЦПТ:
 - Наблюдения в выборке должны быть независимы друг от друга.
 - Большой объем выборки **или** нормальное распределение x

Если σ не известна

Если σ в генеральной совокупности не известна, ее можно оценить по выборочному стандартному отклонению s .

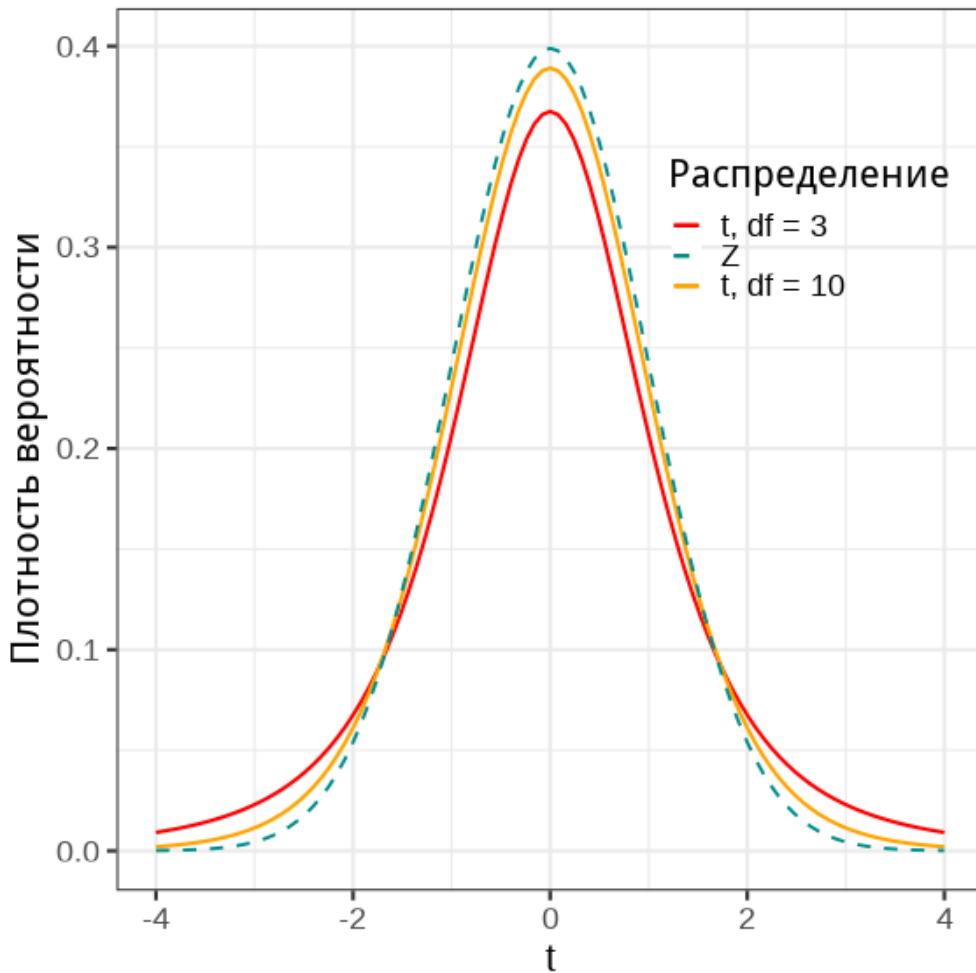
$$\sigma/\sqrt{n} \approx s/\sqrt{n}$$

Тогда, после стандартизации:

$$\frac{\bar{X} - \mu}{SE_{\bar{x}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{df=n-1}$$

То есть, если σ в генеральной совокупности не известна, то стандартизованное распределение выборочных средних подчиняется t -распределению с числом степеней свободы $df = n - 1$

t -распределение

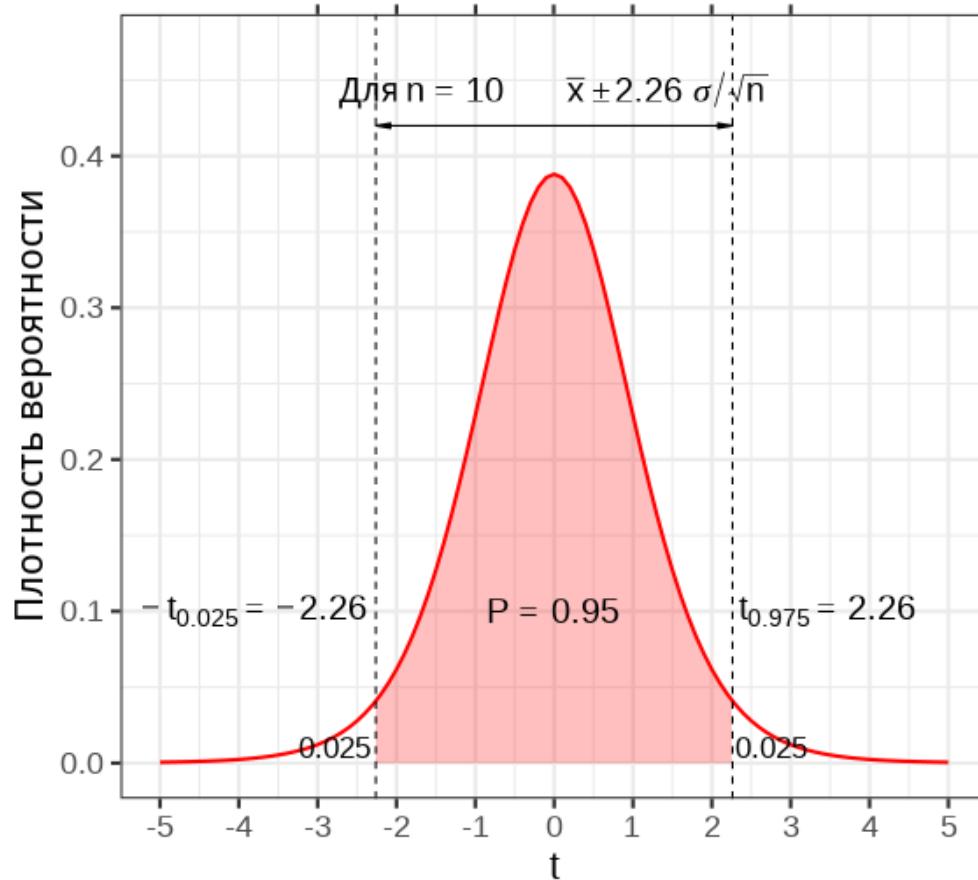


- Симметричное колоколообразное распределение с толстыми хвостами.
- Единственный параметр — число степеней свободы (для доверительного интервала $df = n - 1$).
- При увеличении объема выборки t -распределение приближается к нормальному.

Доверительный интервал из t -распределения

Выборочное распределение средних

$$-\frac{3s}{\sqrt{n}} \leq \bar{x} \leq \frac{3s}{\sqrt{n}}$$



Обязательно используется, если:

- Объем выборки мал.
- σ не известна.

$$\bar{x} \pm t_{\alpha, df} \cdot s / \sqrt{n}$$

$$df = n - 1$$

Условия применимости t -распределения для доверительного интервала

1. ~~Должна быть известна σ в генеральной совокупности.~~
2. Должны выполняться условия, при которых справедлива ЦПТ:
 - Наблюдения в выборке должны быть независимы друг от друга.
 - Большой объем выборки и нет “выбросов” **или** нормальное распределение x

Графическая проверка на соответствие данных нормальному распределению

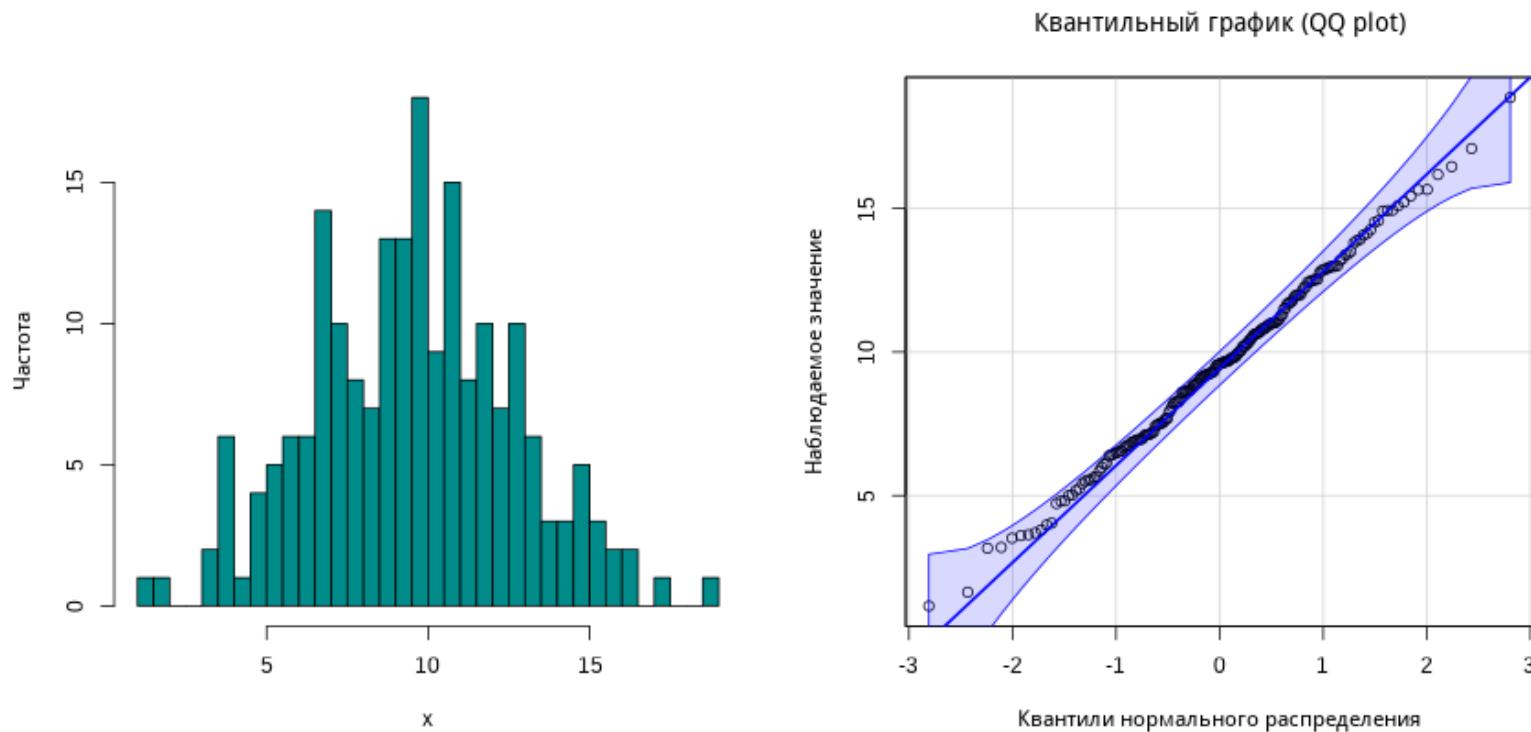
Почему именно графическая проверка?

Не стоит злоупотреблять формальными тестами для проверки на нормальность, т.к. у них тоже есть свои требования, а для больших выборок они склонны быть слишком чувствительными.

В большинстве случаев нам достаточно определить, разумно ли аппроксимировать наблюдаемое распределение нормальным.

Для ответа на этот вопрос вполне подойдет графическая проверка при помощи квантильного графика.

Квантильный график (Quantile-Quantile plot, QQ plot)

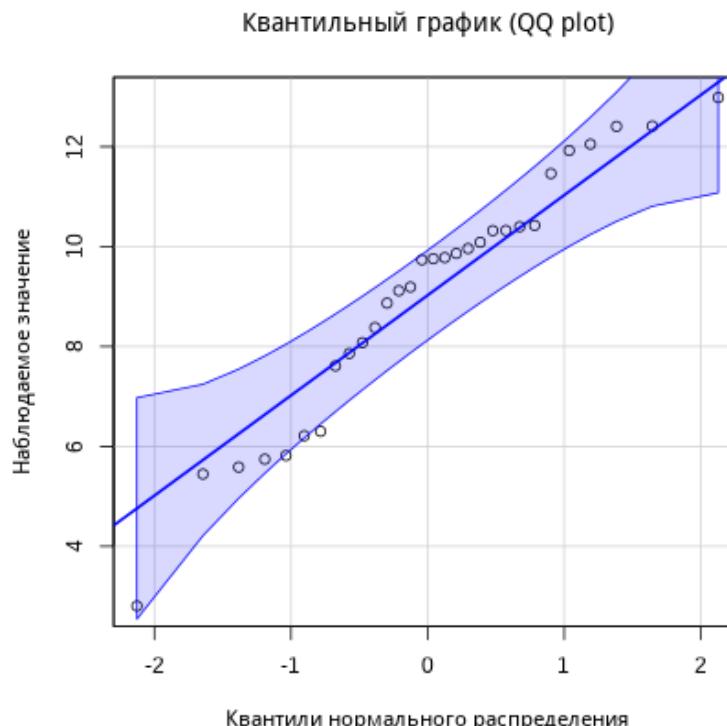
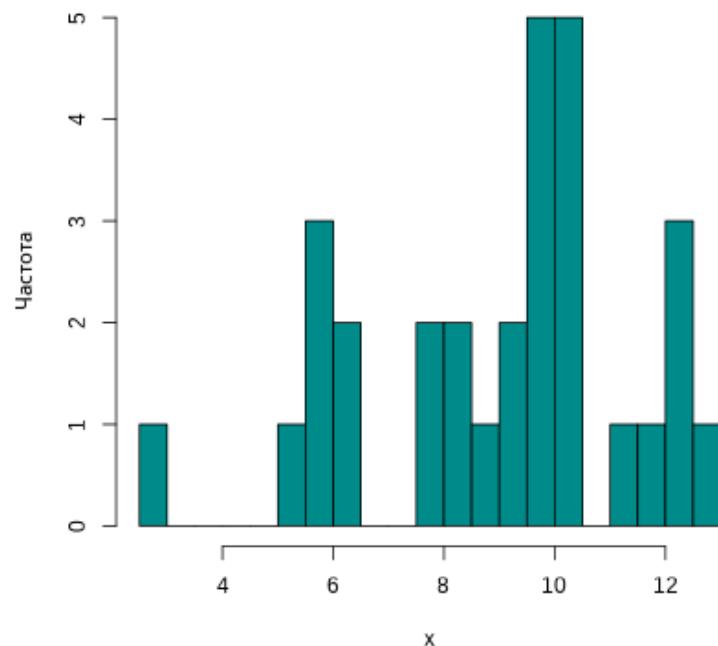


По оси X отложены квантили стандартного нормального распределения.

По оси Y — квантили данных.

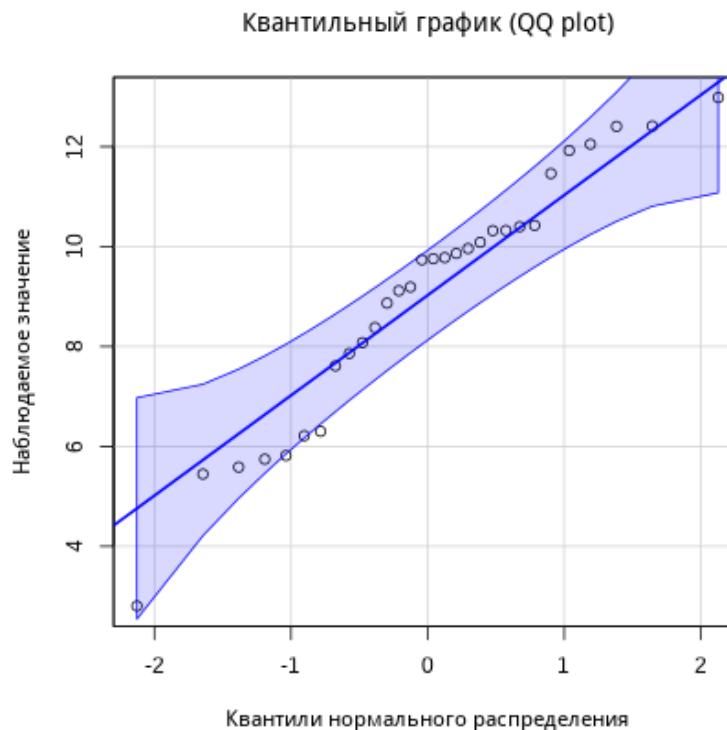
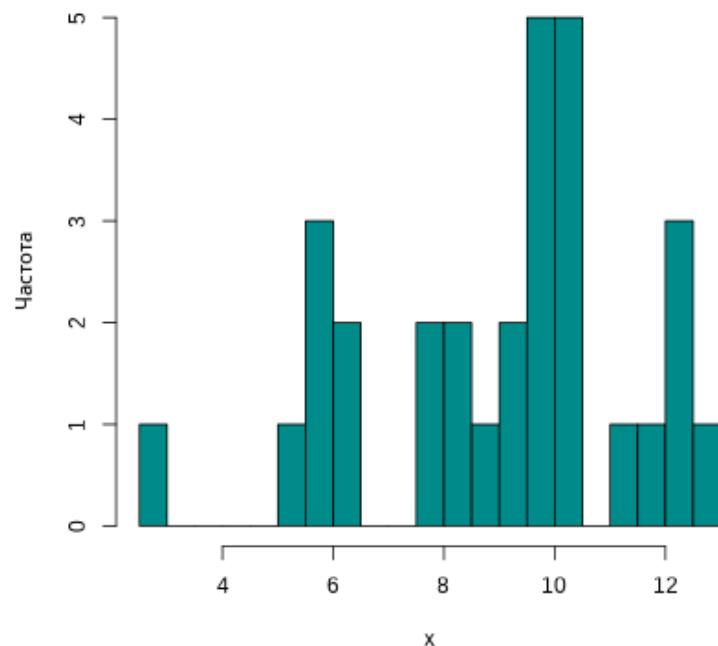
Если x — нормально-распределенная величина, то точки лягут приблизительно по прямой линии.

Квантильный график (Quantile-Quantile plot, QQ plot)



При малых выборках не такая красивая картинка.

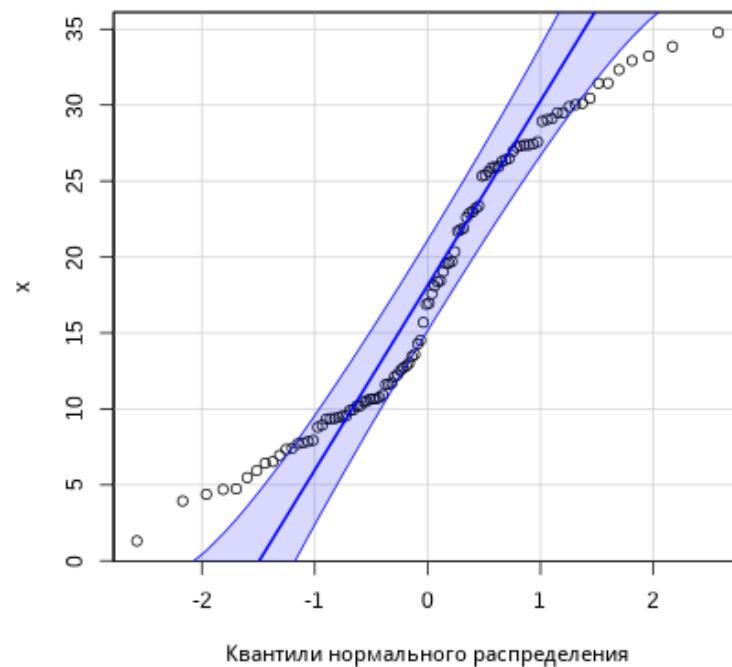
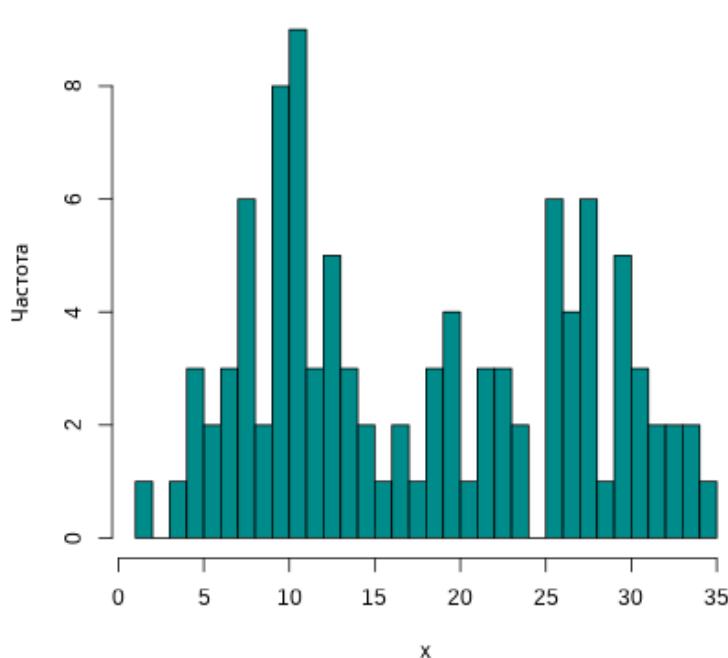
Квантильный график (Quantile-Quantile plot, QQ plot)



При малых выборках не такая красивая картинка.

Как выглядят некоторые отклонения от нормального распределения?

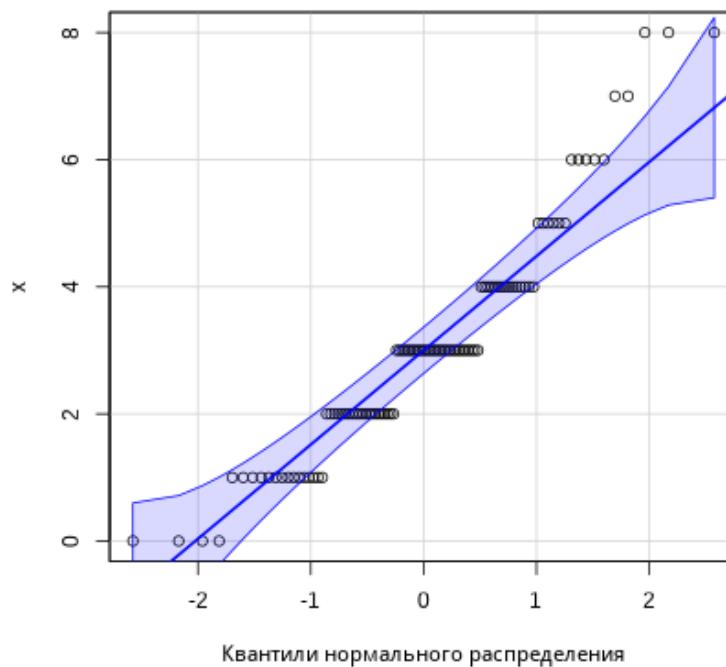
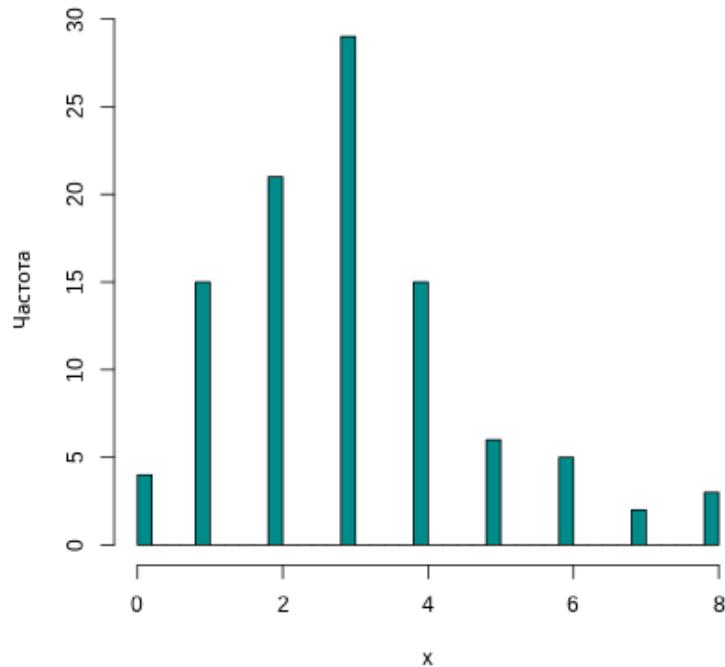
Бимодальное (двувершинное) распределение



Бимодальное распределение — это, в данном случае, смесь двух нормальных распределений.

Квантильный график с перегибом.

Дискретное распределение с длинным правым хвостом



Видно, что x — это дискретная переменная.

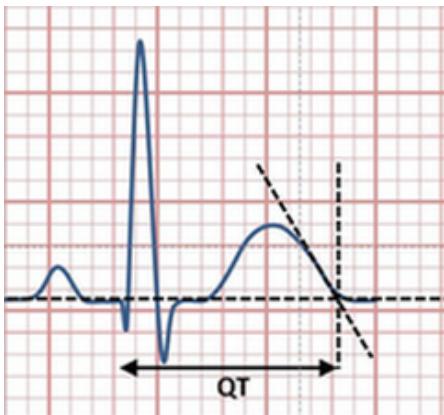
Маленьких значений больше, чем можно ожидать при нормальном распределении.

Длинный правый хвост.

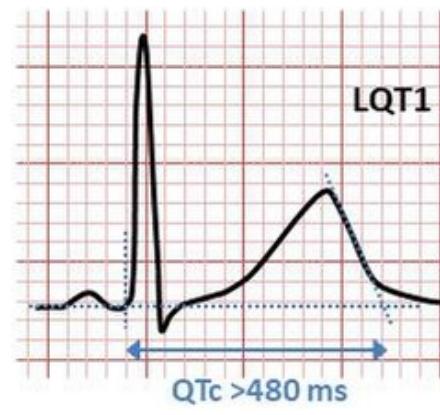
Основы тестирования гипотез

Пример: Интервал QT

Длительность интервала QT на кардиограмме считается предиктором внезапной остановки сердца.



PeaBrainC, CC BY-SA 4.0, via Wikimedia Commons



PeaBrainC, CC BY-SA 4.0, via Wikimedia Commons

Некоторые лекарства способны вызывать удлинение интервала QT.

Нулевая и альтернативная гипотезы

- **Нулевая гипотеза** H_0 чаще всего формулируется как **отсутствие различий** между сравниваемыми объектами.

Например: Использование лекарства не влияет на интервал QT.

Нулевая и альтернативная гипотезы

- **Нулевая гипотеза** H_0 чаще всего формулируется как **отсутствие различий** между сравниваемыми объектами.

Например: Использование лекарства не влияет на интервал QT.

- **Альтернативная гипотеза** H_A формулируется как **присутствие различий**, она обратна нулевой гипотезе, т.е. включает все остальные случаи.

Например: Длительность интервала QT различается у людей использующих и не использующих лекарство.

Нулевая и альтернативная гипотезы

- **Нулевая гипотеза** H_0 чаще всего формулируется как **отсутствие различий** между сравниваемыми объектами.

Например: Использование лекарства не влияет на интервал QT.

- **Альтернативная гипотеза** H_A формулируется как **присутствие различий**, она обратна нулевой гипотезе, т.е. включает все остальные случаи.

Например: Длительность интервала QT различается у людей использующих и не использующих лекарство.

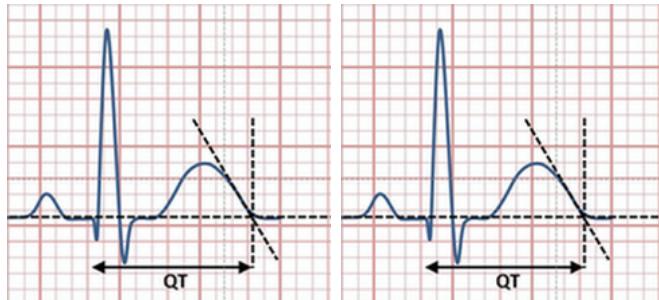
В некоторых случаях допустимо формулировать **двустороннюю альтернативную гипотезу**, т.е. с учетом различий в каком-то одном направлении.

Например: Использование лекарства увеличивает QT.

Нулевая и альтернативная гипотезы – это “два мира”

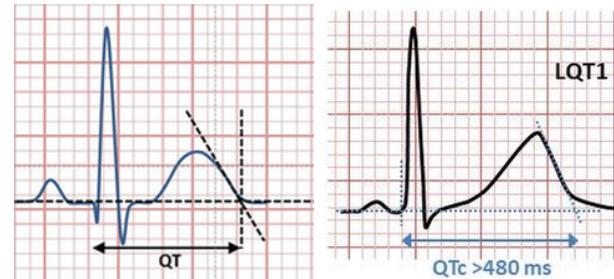
Реальность может находиться в одном из двух состояний:

H_0 верна, длина QT не зависит от применения препарата



PeaBrainC, CC BY-SA 4.0, via Wikimedia Commons

H_0 неверна, QT разный у использующих и не использующих препарат

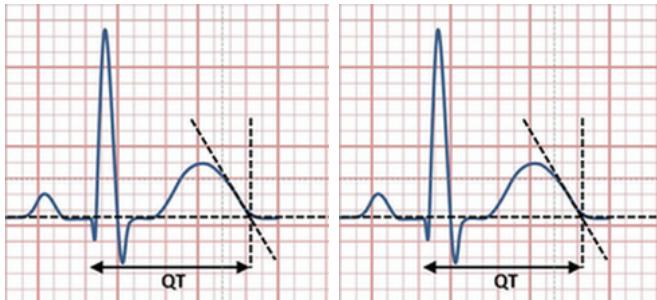


PeaBrainC, CC BY-SA 4.0, via Wikimedia Commons

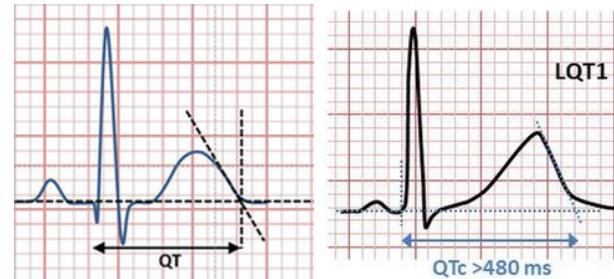
Нулевая и альтернативная гипотезы – это “два мира”

Реальность может находиться в одном из двух состояний:

H_0 верна, длина QT не зависит от применения препарата



H_0 неверна, QT разный у использующих и не использующих препарат



PeaBrainC, CC BY-SA 4.0, via Wikimedia Commons

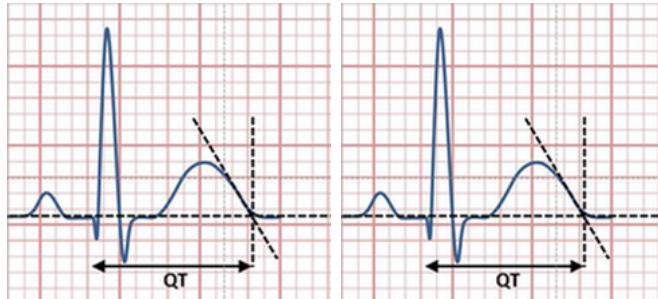
PeaBrainC, CC BY-SA 4.0, via Wikimedia Commons

После статистического теста мы принимаем решение принять или отвергнуть H_0 , но оно не обязательно верно.

Верные и неверные решения

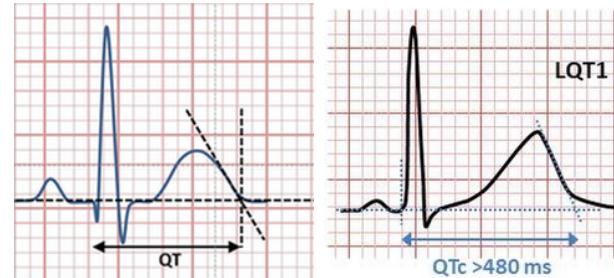
Реальность может находиться в одном из двух состояний:

H_0 верна, длина QT не зависит от применения препарата



PeaBrainC, CC BY-SA 4.0, via Wikimedia Commons

H_0 неверна, QT разный у использующих и не использующих препарат



PeaBrainC, CC BY-SA 4.0, via Wikimedia Commons

После статистического теста возможно 4 исхода:

	H_0 верна	H_0 неверна
Отклонить H_0	Ошибка I рода: нашли то, чего нет, с вероятностью α Ложно-положительный результат	Верно
Сохранить H_0	Верно	Ошибка II рода: не нашли то, что было, с вероятностью β Ложно-отрицательный результат

Тестирование гипотез: Тестовые статистики

1. Формулируем нулевую и альтернативную гипотезы.

Гипотезы выражаются математически в виде тестовых статистик. На этом этапе мы делаем определенные допущения.

Тестирование гипотез: Тестовые статистики

1. Формулируем нулевую и альтернативную гипотезы.

Гипотезы выражаются математически в виде тестовых статистик. На этом этапе мы делаем определенные допущения.

2. Проверяем [условия применимости](#) тестовой статистики.

Тестирование гипотез: Тестовые статистики

1. Формулируем нулевую и альтернативную гипотезы.

Гипотезы выражаются математически в виде тестовых статистик. На этом этапе мы делаем определенные допущения.

2. Проверяем [условия применимости](#) тестовой статистики.

3. По реальным данным вычисляем [эмпирическое значение](#) тестовой статистики.

Тестирование гипотез: Тестовые статистики

1. Формулируем нулевую и альтернативную гипотезы.

Гипотезы выражаются математически в виде тестовых статистик. На этом этапе мы делаем определенные допущения.

2. Проверяем условия применимости тестовой статистики.

3. По реальным данным вычисляем эмпирическое значение тестовой статистики.

Дальше мы должны ответить на вопрос:

Насколько вероятно получить такое или более экстремальное эмпирическое значение, если верна нулевая гипотеза H_0 ?

4. Строим теоретическое распределение тестовой статистики для случая, когда верна H_0 , и оцениваем по нему уровень значимости.

Тестирование гипотез: Тестовые статистики

1. Формулируем нулевую и альтернативную гипотезы.

Гипотезы выражаются математически в виде тестовых статистик. На этом этапе мы делаем определенные допущения.

2. Проверяем условия применимости тестовой статистики.

3. По реальным данным вычисляем эмпирическое значение тестовой статистики.

Дальше мы должны ответить на вопрос:

Насколько вероятно получить такое или более экстремальное эмпирическое значение, если верна нулевая гипотеза H_0 ?

4. Строим теоретическое распределение тестовой статистики для случая, когда верна H_0 , и оцениваем по нему уровень значимости.

5. Решаем сохранить или отвергнуть H_0 .

Увы, мы не сможем узнать, какая гипотеза верна, но поймем, насколько с ней согласуются исходные данные.

Одновыборочный t -тест

Пример для одновыборочного t -теста

Разберемся с одновыборочным t -тестом на вымышленном примере*.



Давайте проверим, есть ли статистически значимые отличия длины интервала QTc после применения лекарства от нормы.

Длительность интервала QT на кардиограмме считается предиктором внезапной остановки сердца. Некоторые лекарства способны вызывать удлинение интервала QT

Нормой считается скорректированный QTc < 450 мс (для мужчин).

После терапии неким лекарством в выборке из 35 мужчин оказалось, что средний QTc $\bar{x} = 463$ мс и $s = 32$ мс.

Одновыборочный t-тест

- $H_0 : \mu = \mu_0$ — Наблюдаемая средняя длина QTc такая же как в норме.
- $H_A : \mu \neq \mu_0$ — Средняя длина QTc отличается от нормы.

μ_0 — это какое-то конкретное значение. В нашей задаче это — 450 мс.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Если выполняется ЦПТ, то одновыборочная t -статистика подчиняется t -распределению с числом степеней свободы $df = n - 1$.

Условия применимости:

- Наблюдения в выборке должны быть независимы друг от друга.
- Объем выборки достаточно велик **или** величины нормально распределены.

Одновыборочный t-тест

- $H_0 : \mu = \mu_0$ — Наблюдаемая средняя длина QTc такая же как в норме.
- $H_A : \mu \neq \mu_0$ — Средняя длина QTc отличается от нормы.

μ_0 — это какое-то конкретное значение. В нашей задаче это — 450 мс.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Если выполняется ЦПТ, то одновыборочная t -статистика подчиняется t -распределению с числом степеней свободы $df = n - 1$.

Условия применимости:

- Наблюдения в выборке должны быть независимы друг от друга.
- Объем выборки достаточно велик **или** величины нормально распределены.

В нашем случае объем выборки достаточно велик, если бы людей было меньше, нужно было бы проверить на нормальность распределения.

Вычислим наблюдаемое значение t -статистики

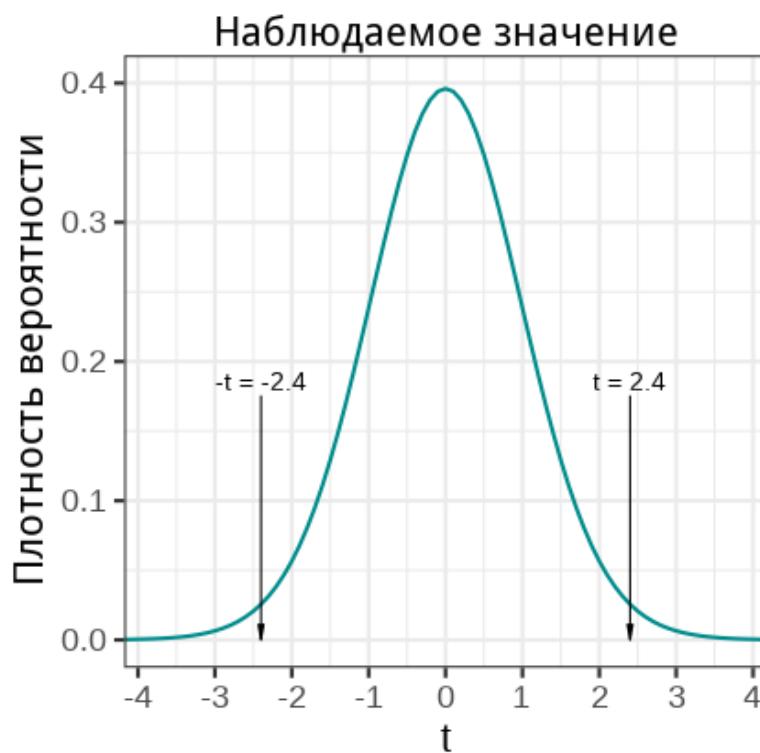
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Наши данные: в норме QTc < 450 мс, а после терапии в выборке из 35 мужчин оказалось, что средний QTc $\bar{x} = 463$ мс и $s = 32$ мс.

Вычислим t -статистику по этим данным:

$$t = \frac{463 - 450}{32/\sqrt{35}} = 2.4$$

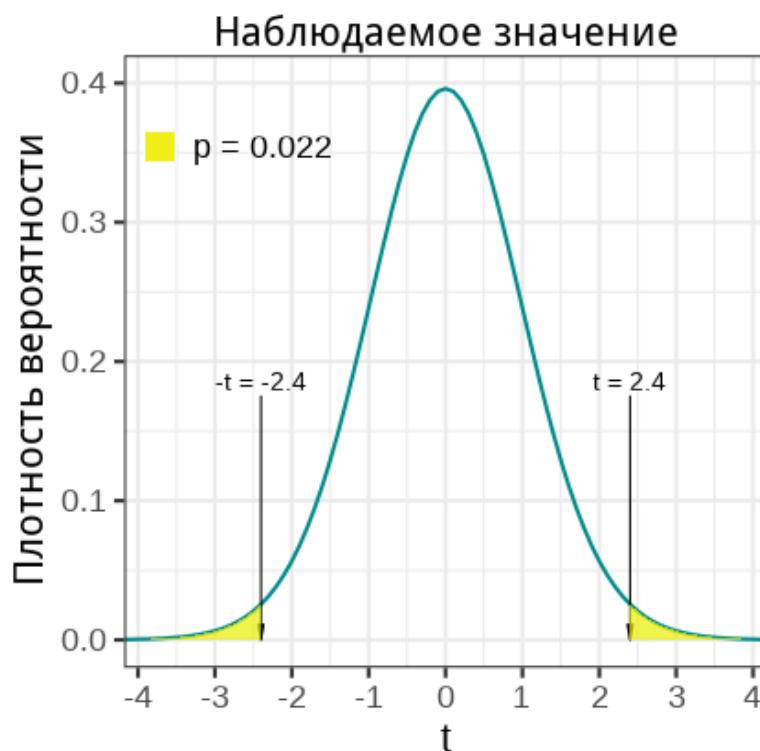
Насколько это значение согласуется с H_0 ?



При H_0 значение t будет близко к нулю.

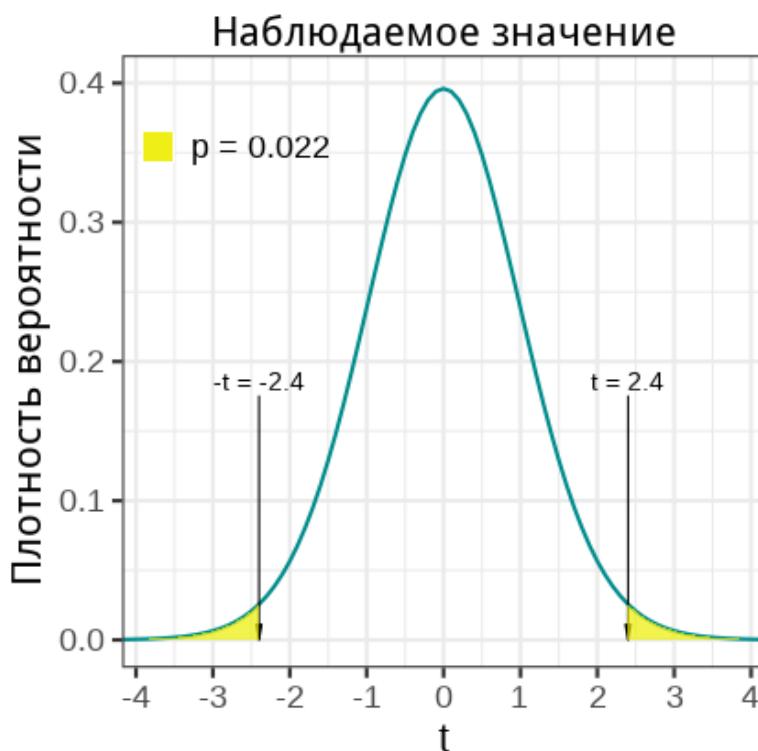
Насколько необычны значения t меньше или больше ± 2.4 ?

Уровень значимости (p -value)



Уровень значимости — это вероятность получить такое же или более экстремальное значение статистики при условии, что H_0 была бы справедлива.
(Определение для двустороннего теста).

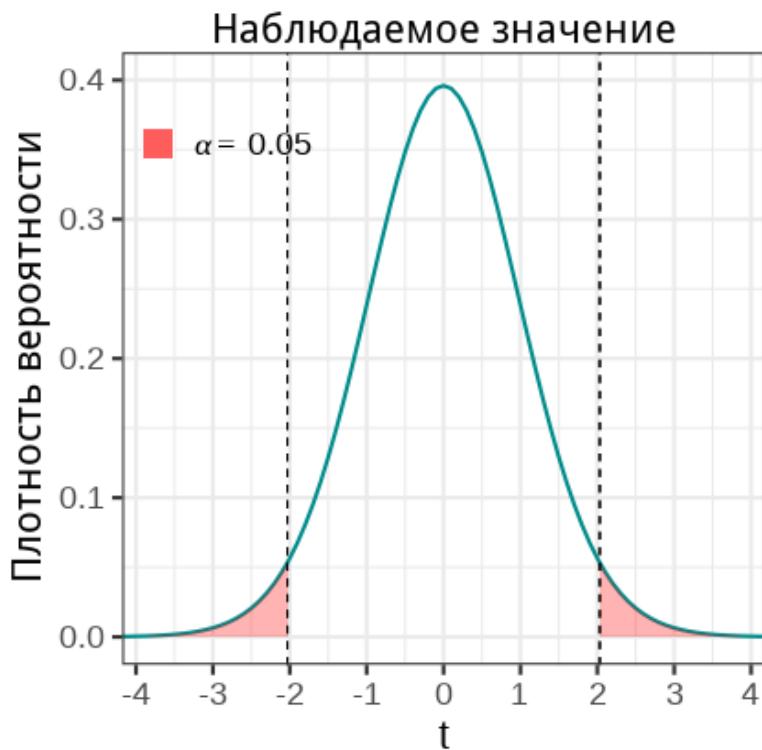
Уровень значимости (p -value)



Можно вычислить значение p

Если бы длительность интервала QTc не отличалось нормы, получить t меньше или больше 2.4 можно было бы с вероятностью $p = 0.022$.

Критический уровень значимости



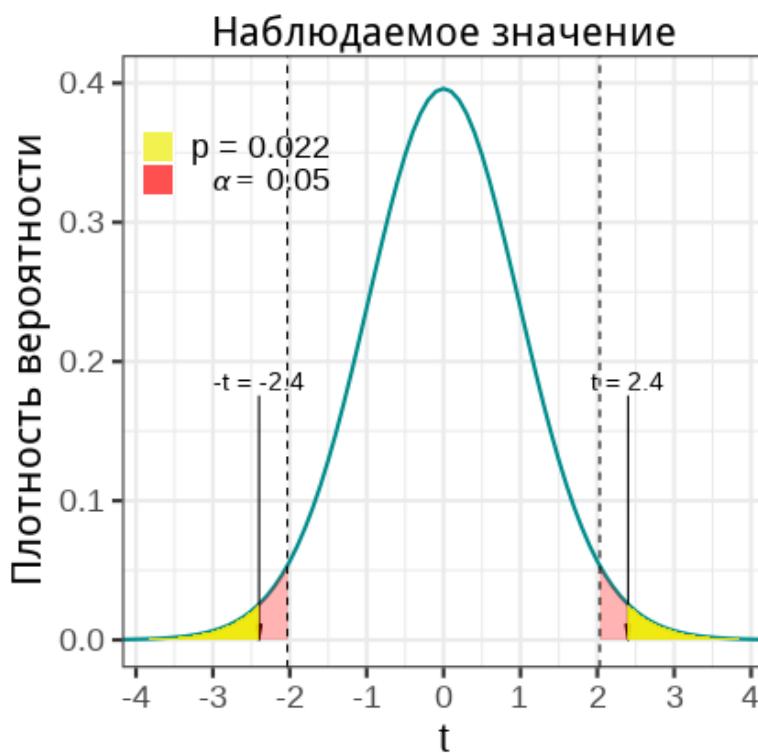
Критический уровень значимости α — это порог для принятия решений.

Обычно используют $\alpha = 0.05$.

Если $p \leq \alpha$ — отвергаем H_0 и принимаем H_A .

Если $p > \alpha$ — сохраняем H_0 , не можем ее отвергнуть.

Принимаем решение



Мы получили $p < \alpha$, поэтому отвергаем H_0 и принимаем H_A .

Фактическая длительность интервала QTc статистически значимо отличается от нормы.

Заблуждения о *p*-values

Заблуждения о *p*-values

Правда ли, что p — вероятность того, что верна сама H_0 ?

Заблуждения о *p*-values

Правда ли, что p — вероятность того, что верна сама H_0 ?

Нет! Значение p всегда считается **при условии, что H_0 верна.**

Но H_0 может быть неверна (вне зависимости от значения p или от нашего решения по поводу судьбы H_0).

Заблуждения о *p*-values

Правда ли, что p — вероятность того, что верна сама H_0 ?

Нет! Значение p всегда считается **при условии, что H_0 верна**.

Но H_0 может быть неверна (вне зависимости от значения p или от нашего решения по поводу судьбы H_0).

Правда ли, что p — это вероятность получить такое значение статистики при справедливой H_0 ?

Заблуждения о *p*-values

Правда ли, что p — вероятность того, что верна сама H_0 ?

Нет! Значение p всегда считается **при условии, что H_0 верна**.

Но H_0 может быть неверна (вне зависимости от значения p или от нашего решения по поводу судьбы H_0).

Правда ли, что p — это вероятность получить такое значение статистики при справедливой H_0 ?

Нет! Вероятность вычисляется как площадь под участком кривой.

Конкретное значение статистики — это точка и под ней нет площади.

Заблуждения о *p*-values

Правда ли, что p — вероятность того, что верна сама H_0 ?

Нет! Значение p всегда считается **при условии, что H_0 верна**.

Но H_0 может быть неверна (вне зависимости от значения p или от нашего решения по поводу судьбы H_0).

Правда ли, что p — это вероятность получить такое значение статистики при справедливой H_0 ?

Нет! Вероятность вычисляется как площадь под участком кривой.

Конкретное значение статистики — это точка и под ней нет площади.

Правда ли, что если $p > 0.05$, то различий между группами на самом деле нет?

Заблуждения о *p*-values

Правда ли, что p — вероятность того, что верна сама H_0 ?

Нет! Значение p всегда считается **при условии, что H_0 верна**.

Но H_0 может быть неверна (вне зависимости от значения p или от нашего решения по поводу судьбы H_0).

Правда ли, что p — это вероятность получить такое значение статистики при справедливой H_0 ?

Нет! Вероятность вычисляется как площадь под участком кривой.

Конкретное значение статистики — это точка и под ней нет площади.

Правда ли, что если $p > 0.05$, то различий между группами на самом деле нет?

Нет! Это всего лишь значит, что наблюдаемый эффект согласуется с нулевой гипотезой.

Но мы могли совершить ошибку II рода, и на самом деле различия вполне могли бы быть.

Статистические ошибки при проверке гипотез

Статистические ошибки при проверке гипотез

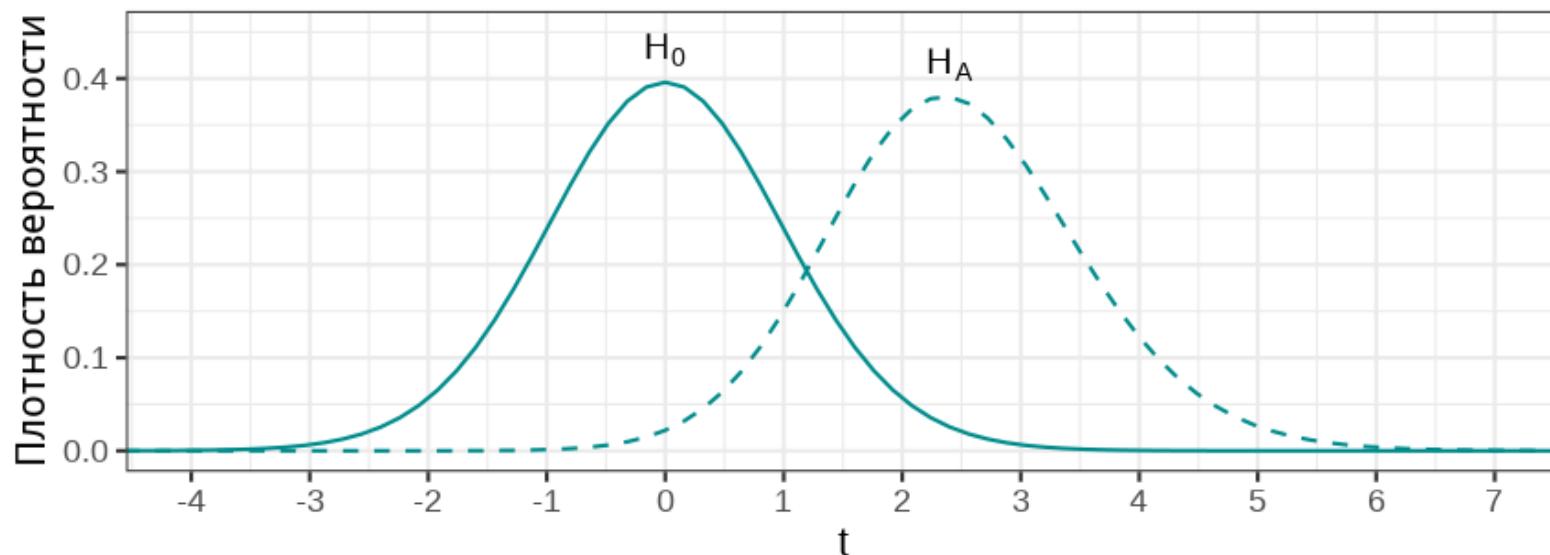
Ошибка I рода: нашли то, чего нет

Ошибка II рода: не нашли то, что было

	H_0 верна	H_0 неверна
Отклонить H_0	Ошибка I рода с вероятностью α Ложно-положительный результат	Верно
Сохранить H_0	Верно	Ошибка II рода с вероятностью β Ложно-отрицательный результат

Как эти ошибки выглядят на графике? Как они взаимосвязаны?

Можно построить теоретические распределения при H_0 и H_A

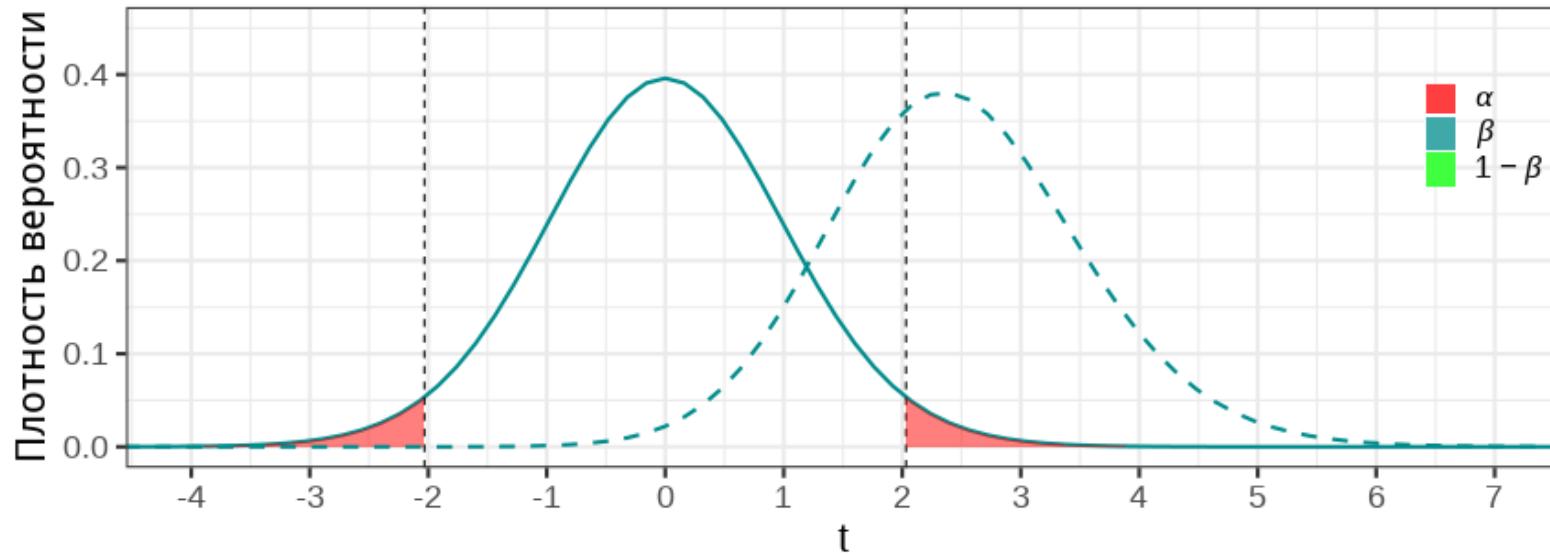


Распределение статистики, когда справедлива H_0 , нам уже знакомо — его мы используем в тестах.

Но может быть справедлива H_A и ее тоже можно описать своим распределением.

При помощи этих распределений можно определить вероятность ошибок различных типов.

Ошибка I рода – найти различия там, где их нет

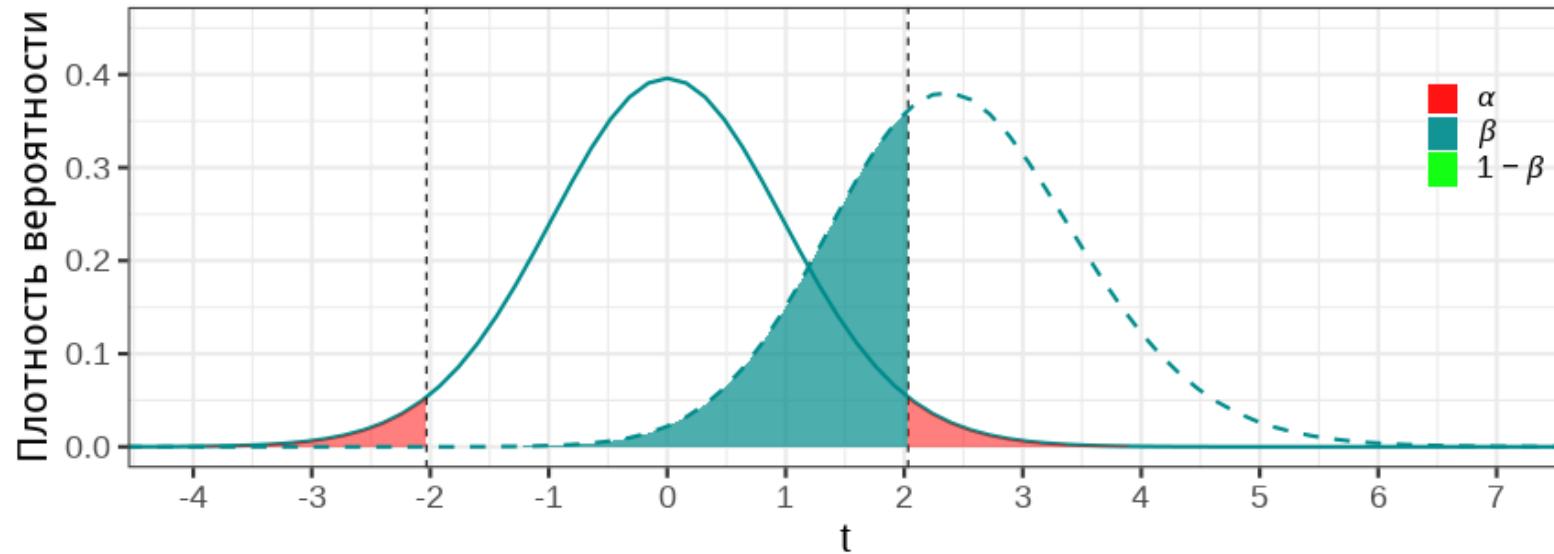


α (критический уровень значимости) — это вероятность ошибки I рода.

Если H_0 справедлива, то при $\alpha = 0.05$ мы отвергаем ее с 5% вероятностью.

Чтобы снизить вероятность таких ошибок, можно уменьшить α .

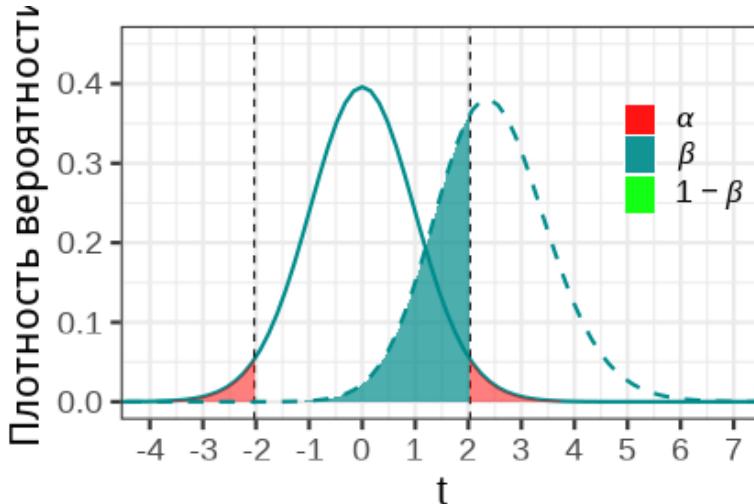
Ошибка II рода – не найти различий, где они есть



β — вероятность ошибки II рода.

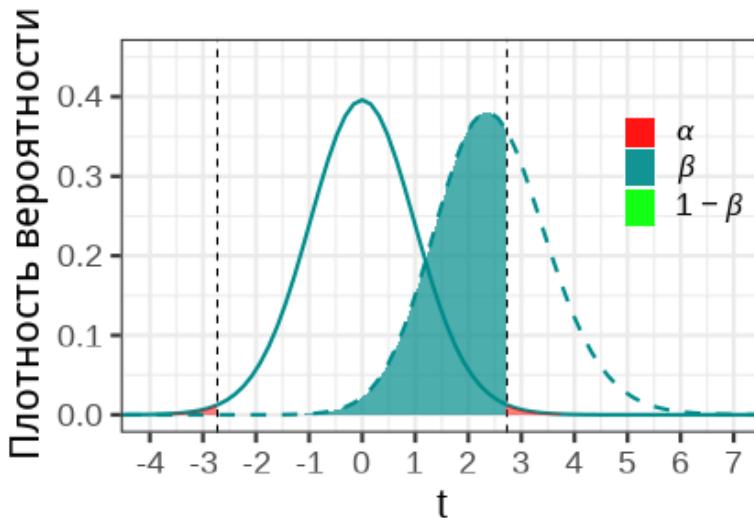
Считается, что допустимо $\beta \leq 0.2$, но часто про нее забывают.

Величина ошибок I и II рода взаимосвязана

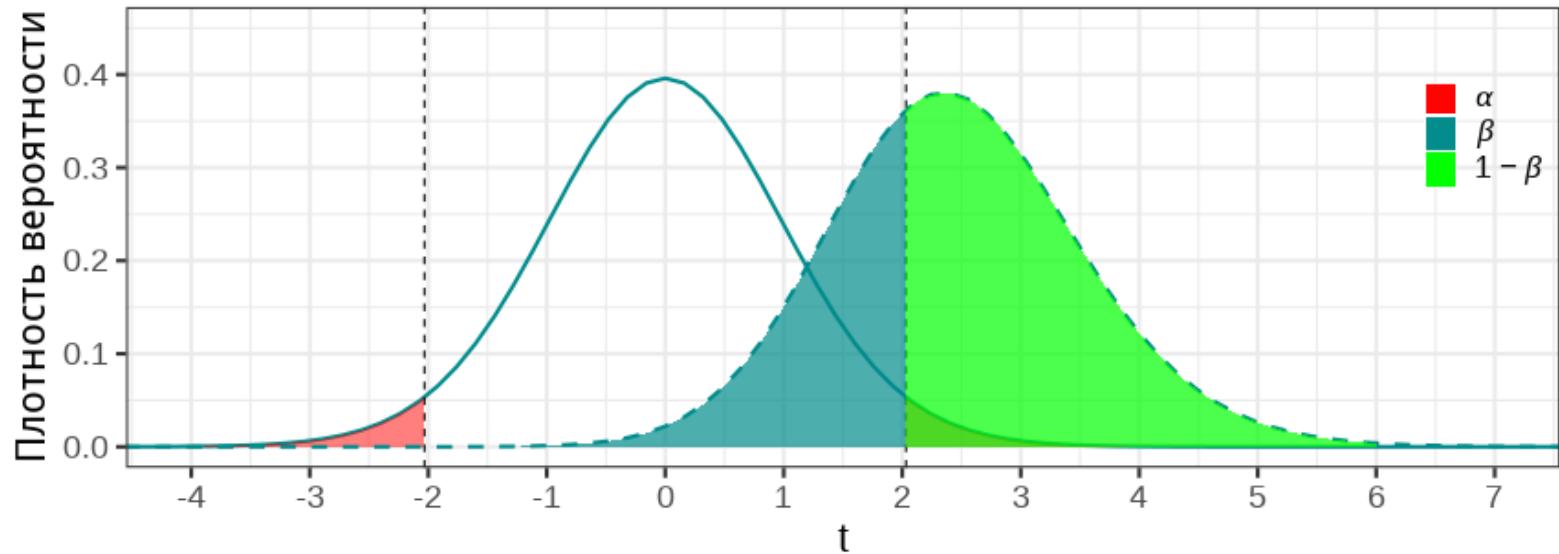


Если мы уменьшаем α (нижний график), то возрастает β .

Т.е. при более жестко выбранном уровне значимости возрастает вероятность не найти различия там, где они есть.



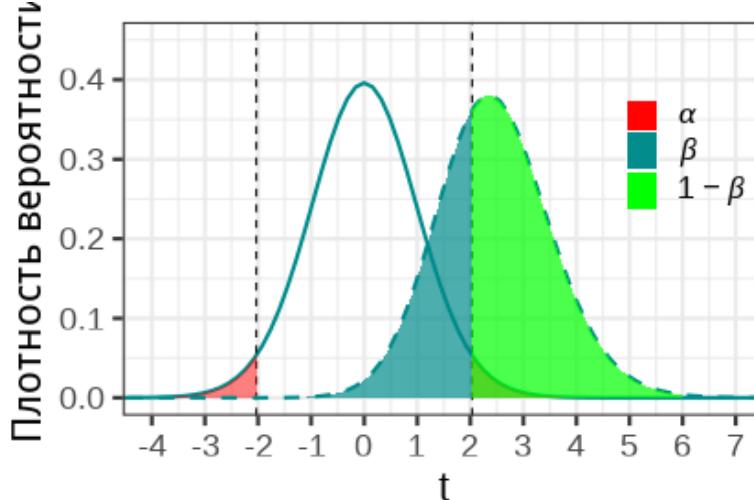
Мощность теста – вероятность найти различия, если они есть



$Power = 1 - \beta$ — мощность теста.

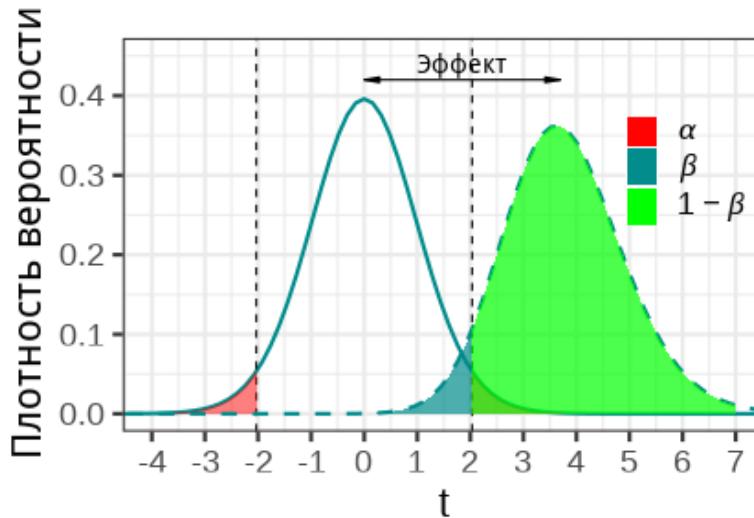
Хорошо, когда мощность не меньше 0.8.

Мощность теста зависит от величины различий



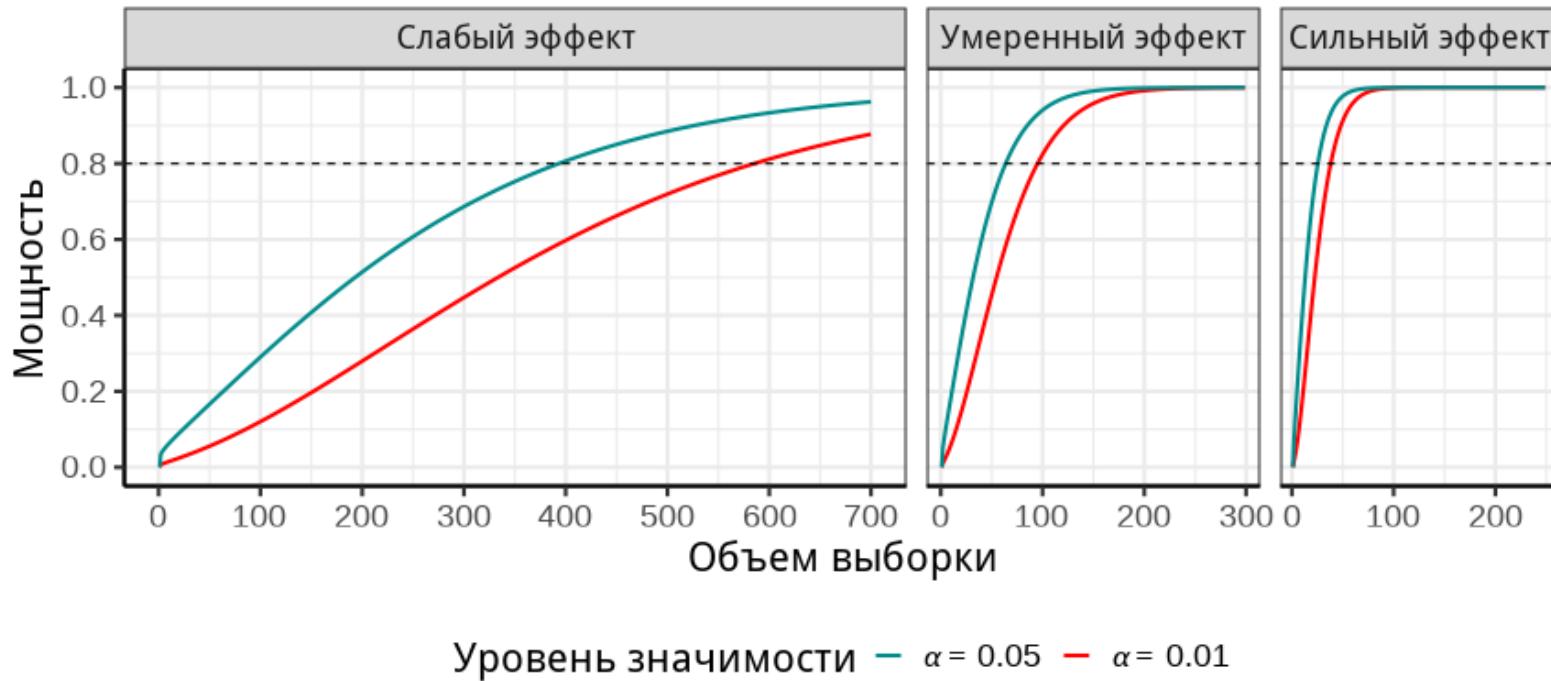
Мощность теста зависит от величины наблюдаемого эффекта (от величины различий).

Если различия велики (график внизу), легче обнаружить их статистическую значимость.



С увеличением объема выборки растет мощность теста

Мощность двухвыборочного t-теста



То есть, если различия очень малы, то хорошей мощности теста 80% сложно достичь, т.к. нужны очень большие объемы выборок.

Summary

Summary:

- Благодаря центральной предельной теореме мы можем строить доверительные интервалы к оценкам различных статистик и тестировать гипотезы.
- Стандартизация — трансформация данных, позволяющая привести их к единой шкале измерений, выраженной в стандартных отклонениях.

Summary:

- Доверительный интервал — это интервал, в который попадает заданный процент выборочных средних.
- Для построения доверительного интервала можно использовать нормальное распределение только тогда, когда известна σ в генеральной совокупности.
- Если же нам приходится оценивать σ по данным как выборочное стандартное отклонение s , то для доверительного интервала будет использоваться t -распределение. Кроме того:
 - Наблюдения в выборке должны быть независимы друг от друга.
 - Большой объем выборки и нет “выбросов” **или** нормальное распределение x

Summary:

- При тестировании гипотез формулируются две гипотезы: нулевая и альтернативная (двусторонняя или односторонняя).
- Уровень значимости — это вероятность получить такое же или более экстремальное значение статистики при условии, что H_0 была бы справедлива.
- При тестировании гипотез возможны ошибки I (ложноположительный результат) или II рода (ложноотрицательный результат).
- Мощность теста — это способность выявлять различия, когда они есть. Она зависит от объема выборки и выбранного порогового уровня значимости для принятия решений по поводу H_0 .

ЧТО ПОЧИТАТЬ

- Quinn, G.G.P., Keough, M.J., 2002. Experimental design and data analysis for biologists. Cambridge University Press.
- Sokal, R.R., Rohlf, F.J., 1995. Biometry (3rd edn). WH Freeman and company: New York.
- Zar, J.H., 2010. Biostatistical Analysis. Prentice Hall: New York.