

**Выборка** - это модель ГС, максимально отражающая её свойства.

**Стратифицированная выборка** - ГС разбивается на похожие группы (страты), а потом выбираются представители из разных страт.

В кластерной выборки, похожие объекты делятся на группы (районы города) и исследуются несколько.

Медиана ГС обозначается как М.

**Унимодальная выборка** – это выборка, имеющая одну моду.

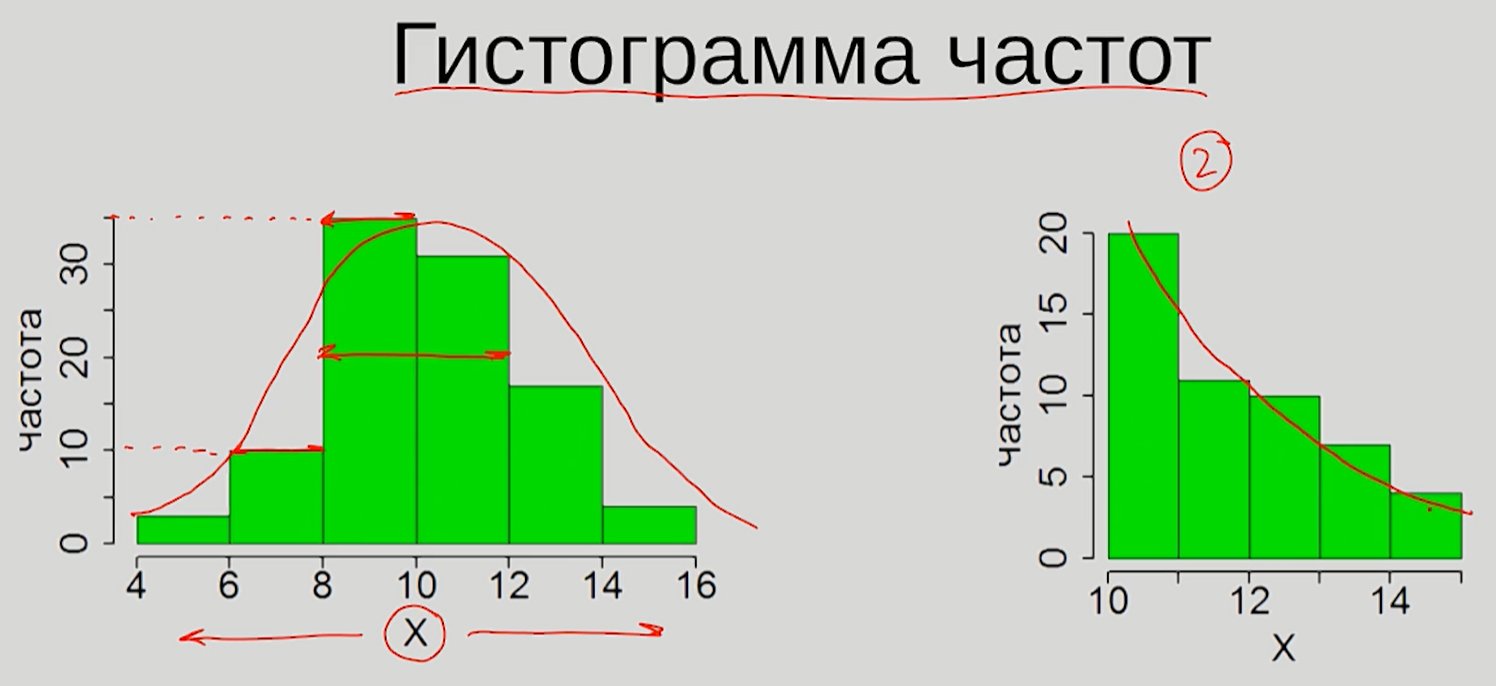
Если выборка полимодальная или имеет явные выбросы (ненормальное распределение), то лучше использовать медиану или моду.

Типы переменных: количественные (дискретные и непрерывные), номинативные (порядковые, бинарные, ранговые). Номинативные переменные имеют числовое значение, которое не позволяет произвести арифметические операции.



Ранговые переменные возможно лишь сравнить, но нельзя определить во сколько раз и на сколько они отличаются, если явно об этом не сообщается в другом признаке.

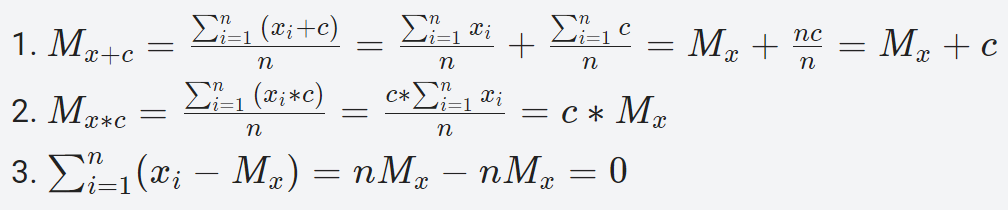
**Гистограмма частот** – сколько раз встречается наблюдение с признаком X в диапазоне.







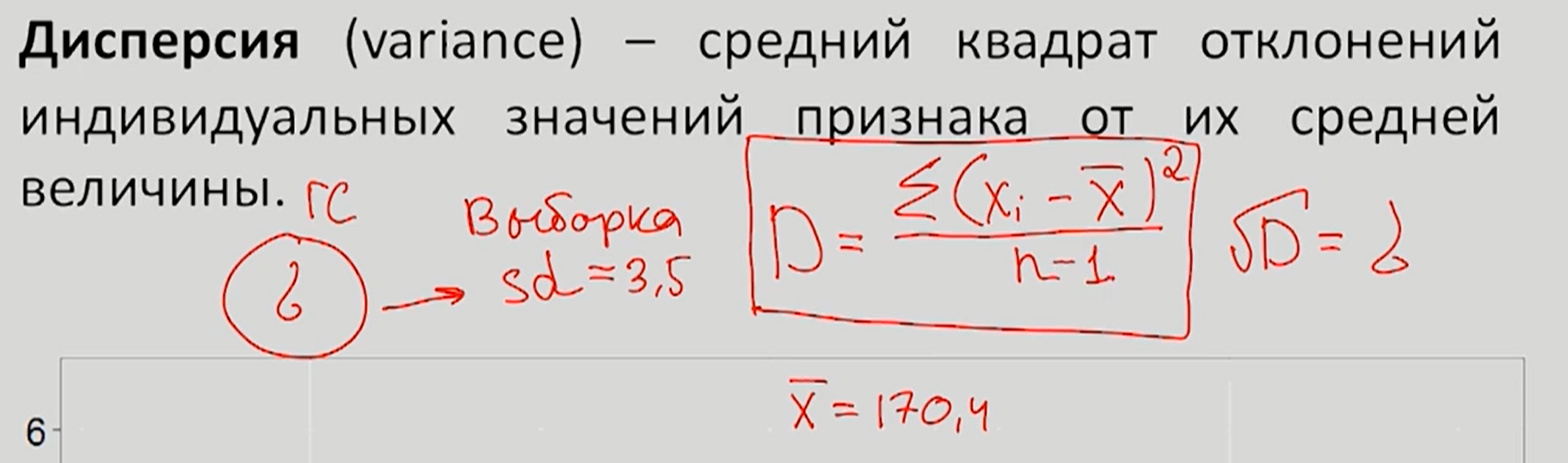
Если к каждому значению выборки (ГС) прибавить константу, тогда среднее значение этой выборки (ГС) будет на константу выше исходного. То же самое работает и при произведении на константу.



**Размах** – разница между максимальным и минимальным значением. Гораздо точнее считать отклонение каждого значения от среднего.

**Дисперсия** – это средний квадрат отклонения конкретного признака от среднего значения признака; мера среднего отклонения значения.

Т.к. дисперсия увеличивает единицу измерений к квадратным, то существует СКО. **Среднеквадратичное отклонение** ГС (у выборки средневыборочное отклонение) позволяет вернуться к исходной единице измерения и показывает реальный размер отклонения значений от среднего.



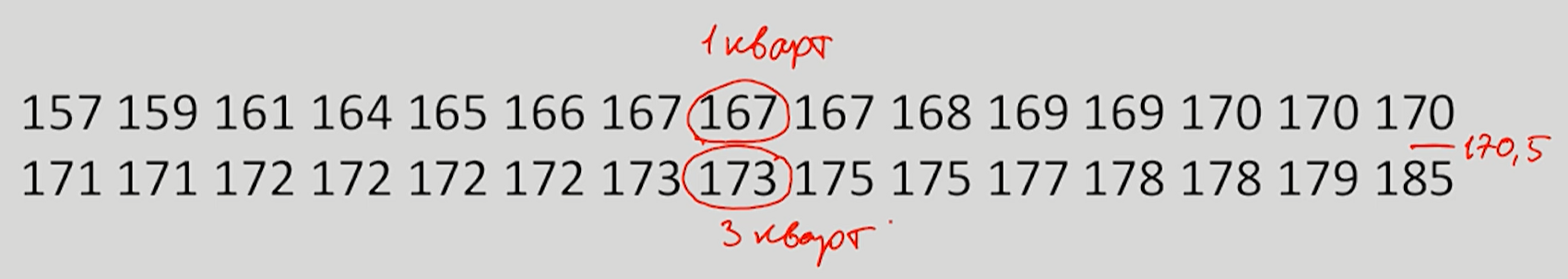
В выборочной дисперсии в знаменателе n – 1, а в ГС дисперсии только n. Это связано со степенями свободы.

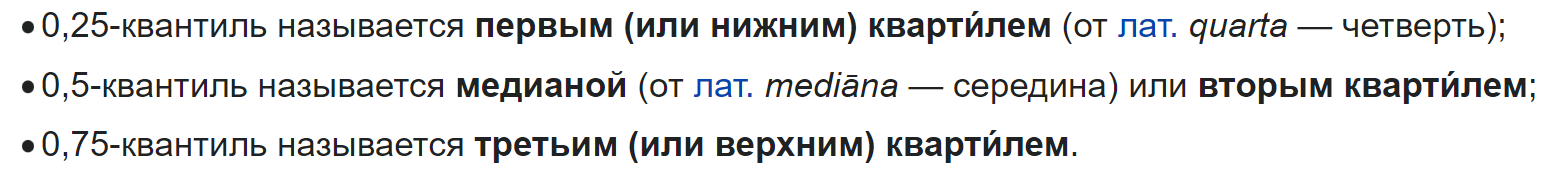


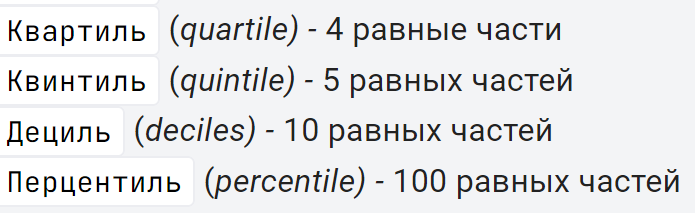
Может ли показатель стандартного отклонения принимать отрицательные значения?

Не может, стандартное отклонение всегда неотрицательное.

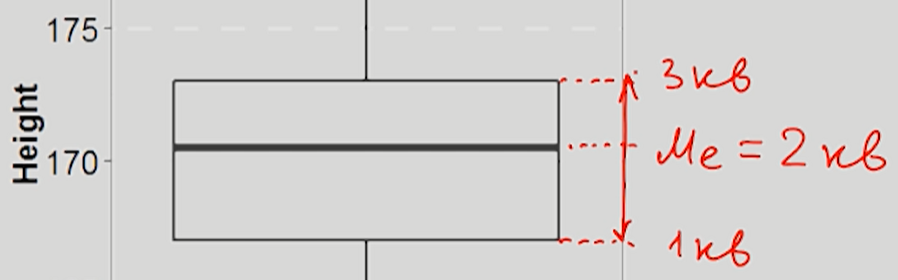
Квантиль – это n - 1 точки, которые делят упорядоченную выборку на n равных частей. Квартиль – это три точки, делящие упорядоченную выборку на 4 равные части.



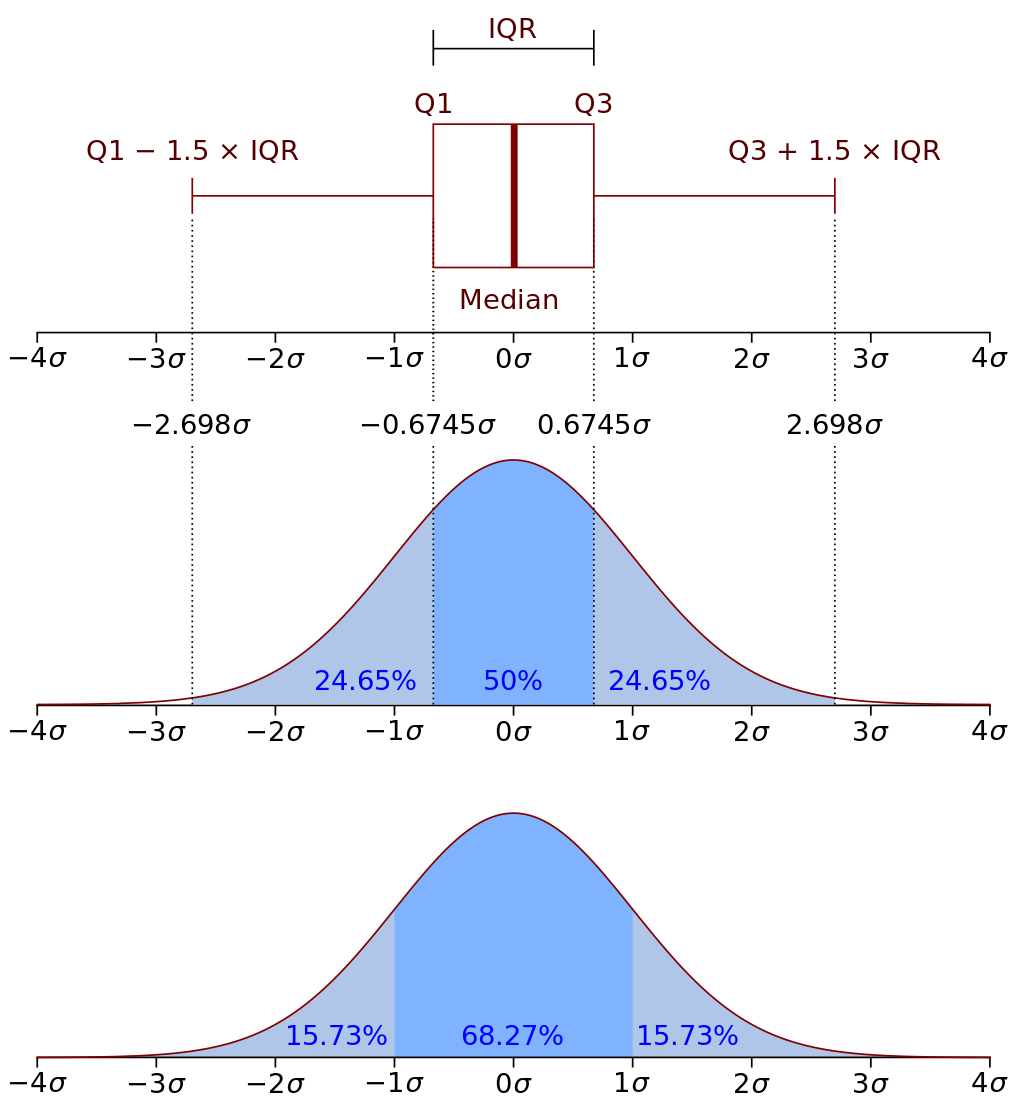




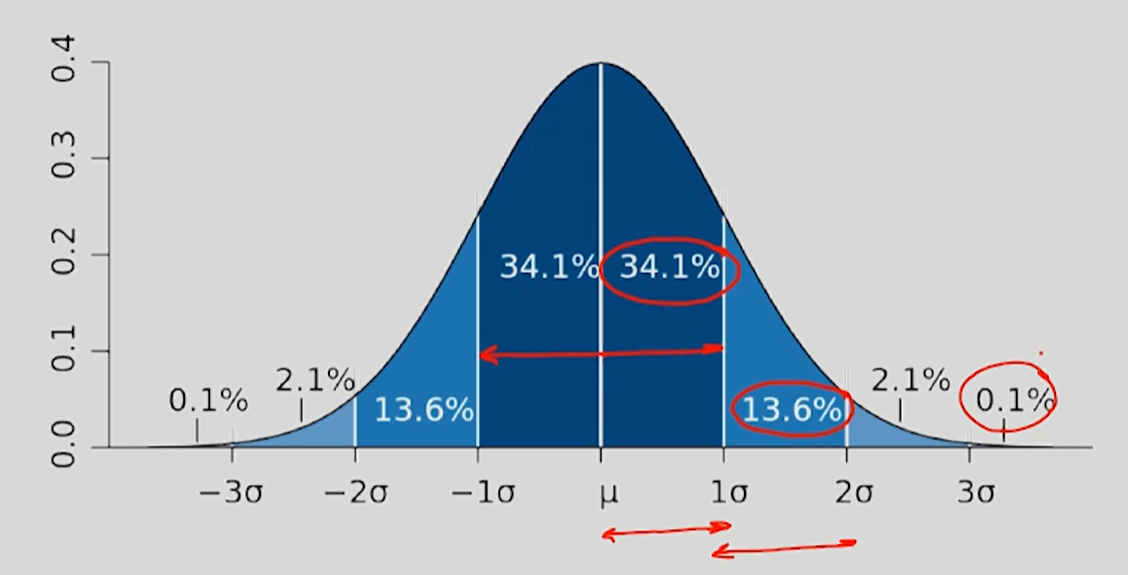
Межквартельный размах – это разность между третьим и первым квартилем, чем эта разность выше, тем выше вариативность исследуемого признака. Ниже изображен Box Plot (ящик с усами).



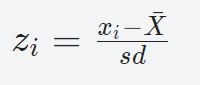
Границы усов находятся в 1.5 \* Межквартельный размах от 1 и 3 их квартилей, те значения, которые находятся выше или ниже, отмечаются отдельно кружочками (выбросы). Box plot призван наглядно продемонстрировать отличие двух и более выборок друг от друга.



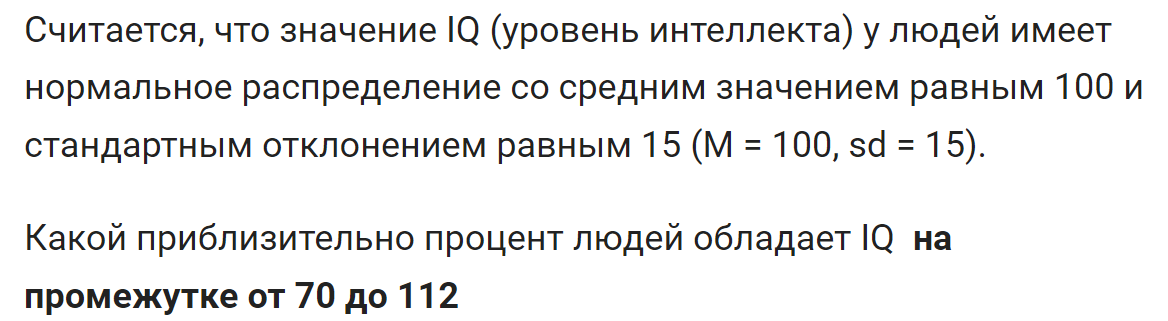
Нормальное распределение



Z – преобразование – это перевод системы координат признака X в такую систему, где X-среднее будет равно 0, а значение 1 (M = 0, sd = 1) будет иметь расстояние от среднего до СКО.



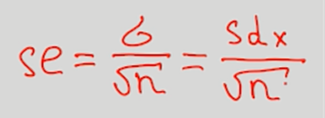
Калькулятор распределений: <https://gallery.shinyapps.io/dist_calc/>



Находим табличные значения для 112 и отнимаем от табличного значения для 70. Затем округляем полученный результат до целых. Единица измерений %.

При нескольких выборках ГС, распределение средних значений выборок будет уже если, количество наблюдений каждой выборки возрастет. Чем больше размер выборки, тем больше она напоминает (приближается к) ГС. Среднее от распределения средних по выборке будет очень близко от среднего значения ГС.

**Стандартная ошибка среднего** (СОС) – это СКО от распределения средних выборочных. Она равна отношению стандартного отклонения на корень из числа наблюдений. Чем больше наблюдений, тем меньше стандартная ошибка среднего. СОС позволяет лишь на одной выборке продемонстрировать поведение ГС и других выборок (если число наблюдений более 30 и текущая выборка репрезентативна).



СКО выборки называется **стандартным отклонением**.

Калькулятор центральной предельной теоремы средних <https://gallery.shinyapps.io/CLT_mean/>

Как соотносятся стандартная ошибка среднего и выборочное стандартное отклонение исследуемого признака (при размере выборки n > 1)?

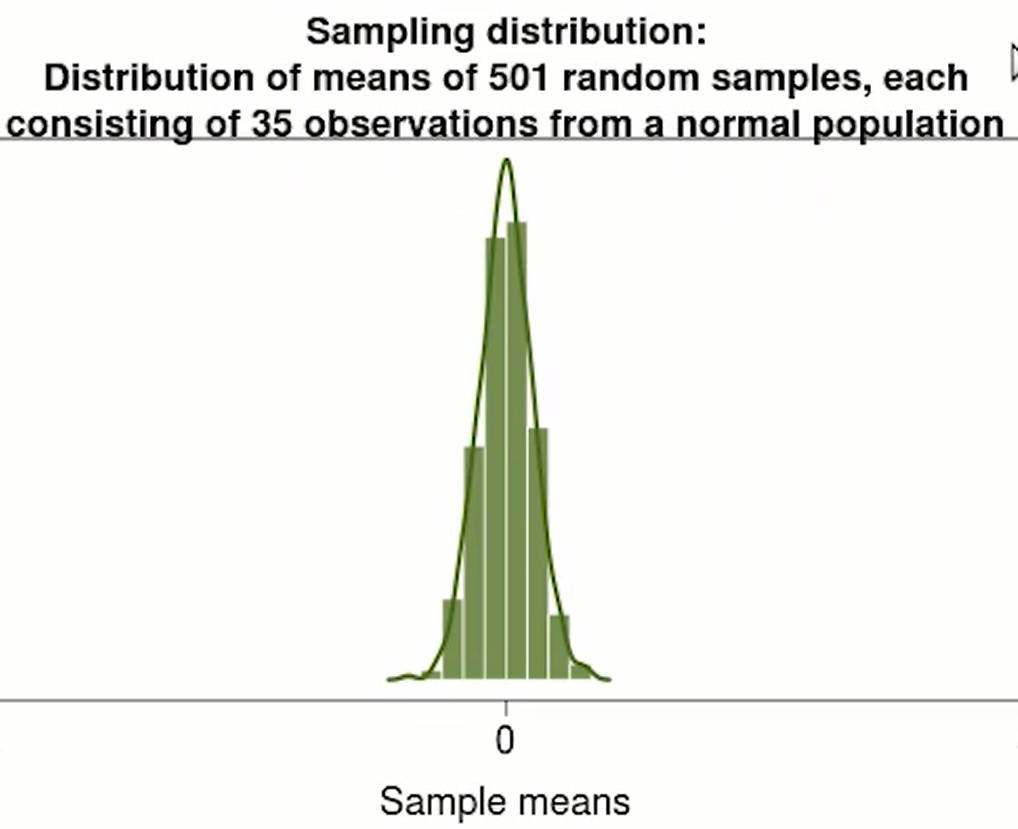
Стандартная ошибка всегда меньше, чем стандартное отклонение

Распределение выборочных средних является нормальным, со средним равным среднему значению признака в генеральной совокупности.

Стандартная ошибка среднего тем меньше, чем больше объем выборки и меньше вариативность исследуемого признака.

**Стандартная ошибка среднего** - это среднеквадратическое отклонение распределения выборочных средних.

Чем меньше стандартная ошибка среднего, тем реже выборочные средние будут сильно отклоняться от среднего в генеральной совокупности.



**Доверительный интервал** (измеряемый в СКО) – это промежуток относительно которого с некоторой %-вероятностью можно быть уверенным, что среднее наблюдаемое значение признака выборки располагается в данном интервале.

95 % всех наблюдений (выборочных средних) лежит в диапазоне приблизительно равном +/- 2 стандартных отклонений от среднего ГС, а если быть максимально точным, то +/-1.96 \* СКО.

При расчете доверительного интервала для каждого из выборочных средних с определенной вероятностью (95%) среднее ГС войдёт в +/- (1.96) СКО от выборочного среднего.

Имея на руках только одну выборку, мы можем сделать вывод о распределении всех выборочных средних (**центральная предельная теорема**). В другой выборке мы получили бы другое значение sd, и в итоге другой доверительный интервал. Но в этом и смысл: в 95 % выборок рассчитанный доверительный интервал включал бы в себя среднее генеральной совокупности.

Если мы рассчитали 95% доверительный интервал для среднего значения, то какие из следующих утверждений являются верными?

Мы можем быть на 95% уверены, что среднее значение в генеральной совокупности принадлежит рассчитанному доверительному интервалу.

Если многократно повторять эксперимент, для каждой выборки рассчитывать свой доверительный интервал, то в 95 % случаев истинное среднее будет находиться внутри доверительного интервала.

Доверительные интервалы используются для оценки параметров ГС, основываясь лишь на выборочных данных.

Если у кого возник вопрос, почему 1,96 и 2,58, то простой ответ - из таблиц. Нет, 1,96 это не просто "неокруглённые" 2 сигмы, и 2,58 это не "неокруглённые" 3 сигмы. Т.е. надо запомнить:

1) для расчета 95% доверительного интервала используйте 1.96, (Z-норм = 1.96)

2) для расчета 99% доверительного интервала используйте 2.58 (Z-норм = 2.58)

Если бы в нашем примере мы увеличили объем выборки в два раза (при условии, что показатель стандартного отклонения остался неизменным), то 95% доверительный интервал

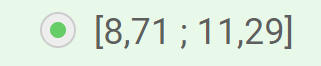
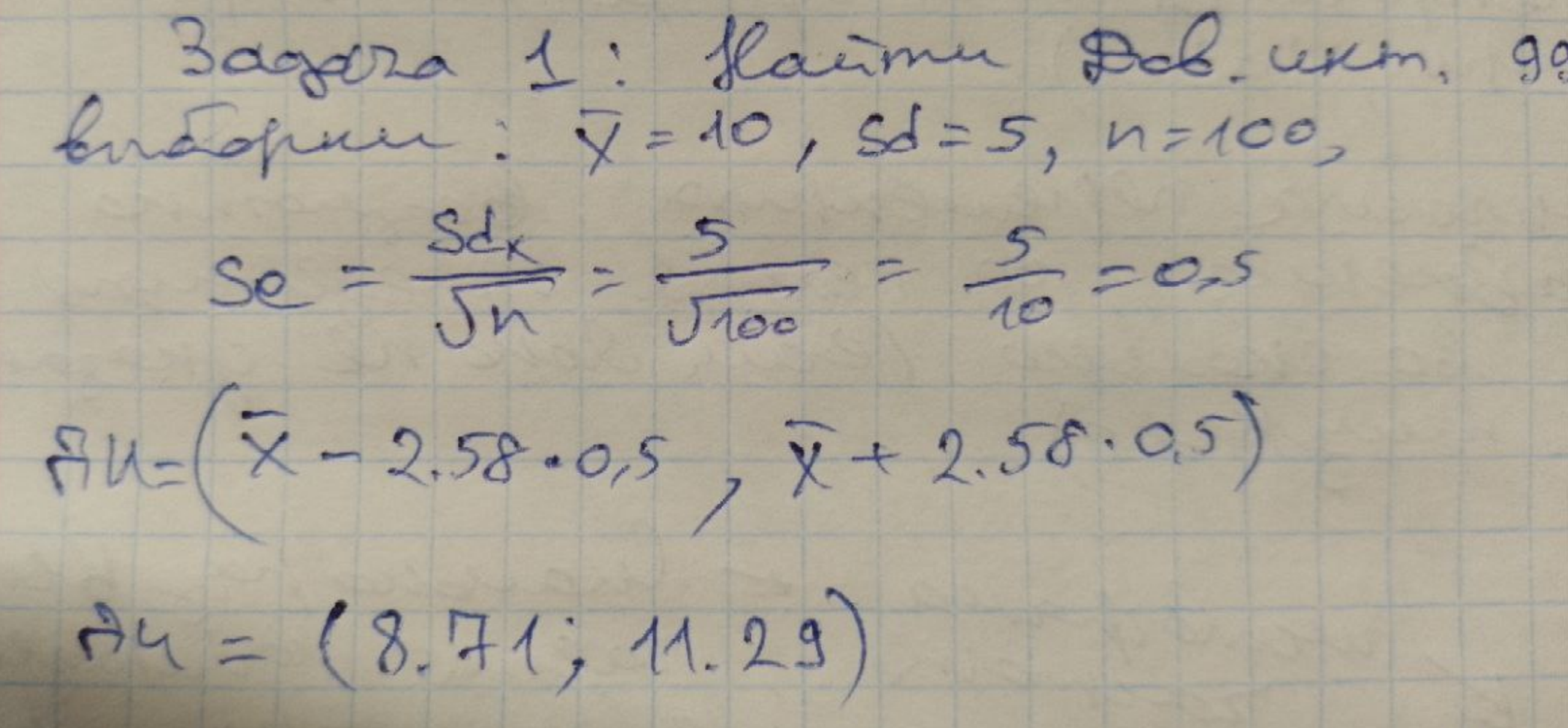
стал бы более узким

Все дело в том, что уменьшился бы показатель СОС, а значит к выборочному среднему меньше бы прибавилось справа и меньше бы отнялось слева.

В центре 95% доверительного интервала, рассчитанного по выборочным значениям, находится:

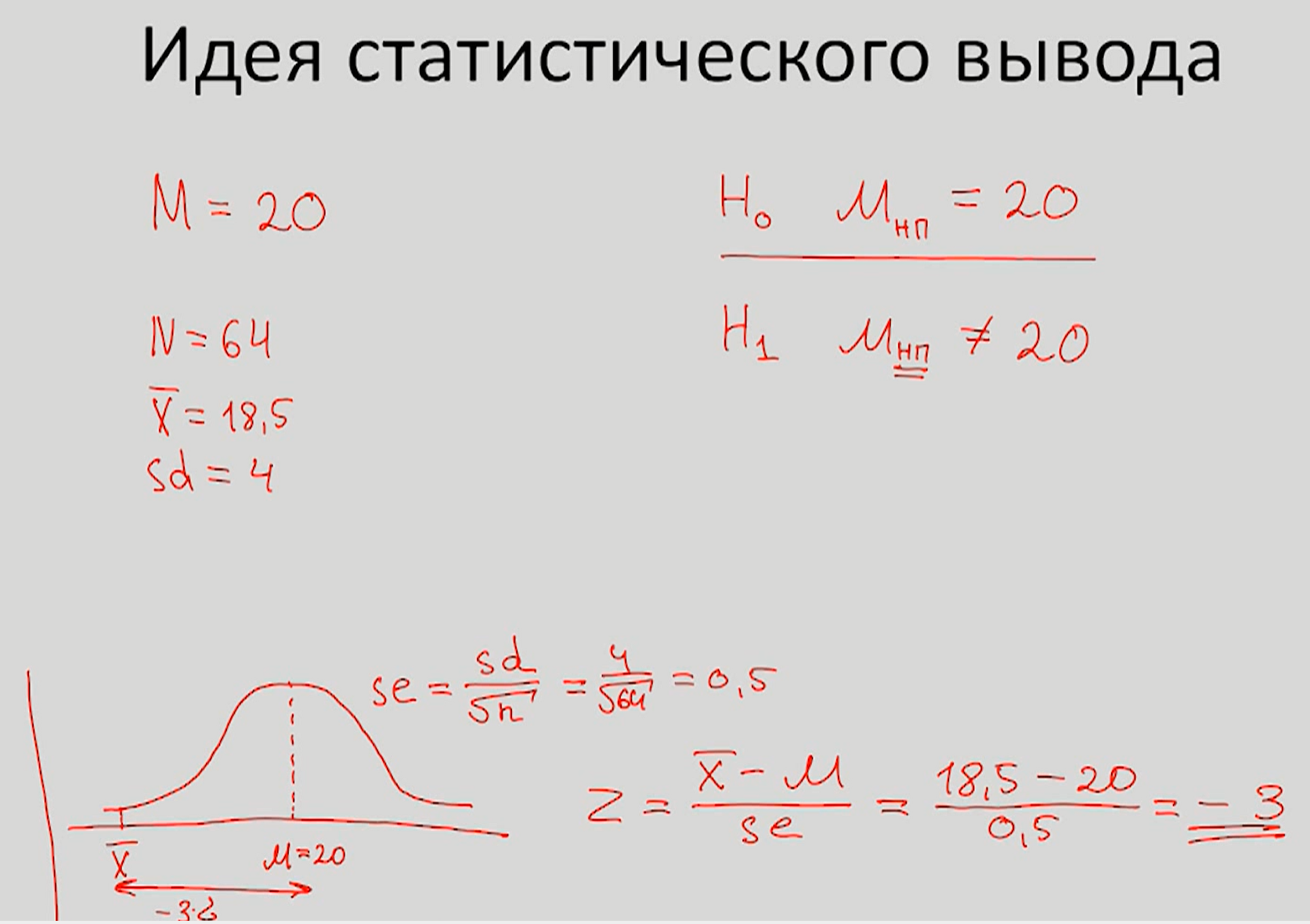
Выборочное среднее значение

Рассчитайте 99% доверительный интервал для следующего примера:

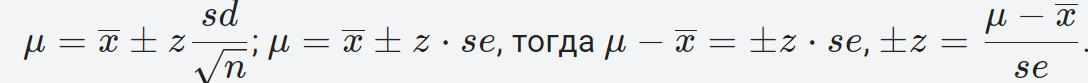


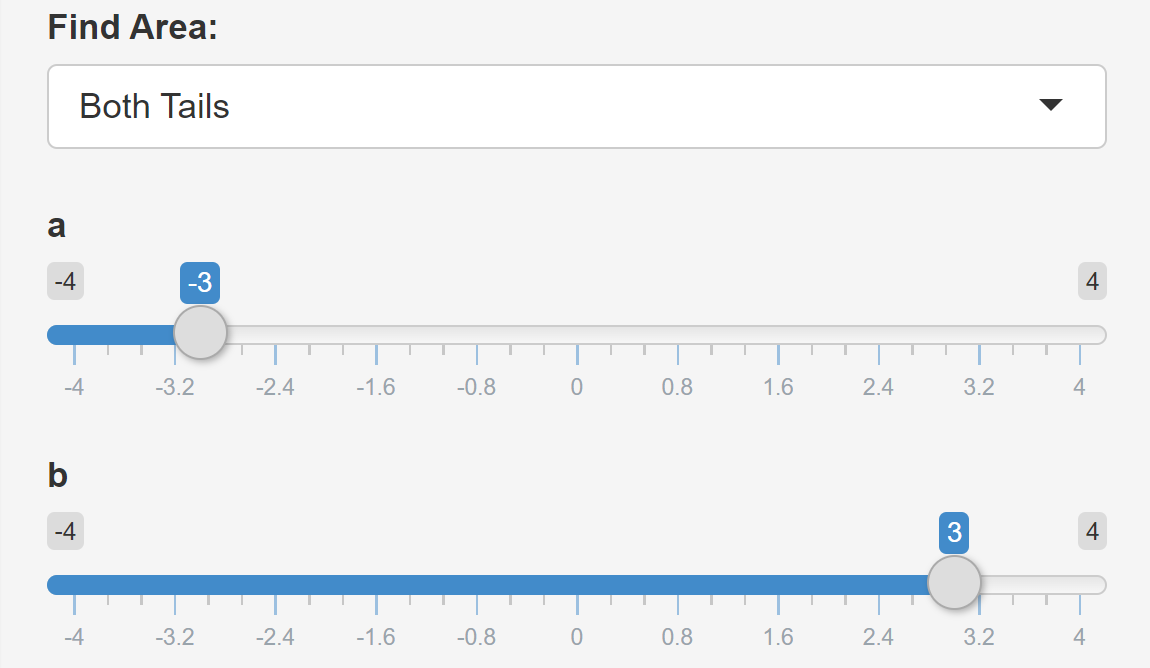
Если я правильно понимаю, то обычно Н0 гипотеза предполагает отсутствие результат (неблагоприятный исход), а Н1 гипотеза предполагает результат (благоприятный исход).

**Стандартная ошибка среднего (СОС) – среднеквадратичное отклонение ГС???**



При среднего ГС (М) равного 20, среднее выборочное значение отклонилось от него на -3 сигмы влево.





Рассчитав на калькуляторе <https://gallery.shinyapps.io/dist_calc/> (при среднем 0 и СКО 1) и установив а на -3 и б на +3, мы выясним, что только около 0.27% выборки падает за пределы от а до б.

Тогда возвращаясь к примеру, где выборочное среднее 18.5 получим, что вероятность получить такое выборочное среднее менее 1%.

Всегда в самом начале мы допускаем H0 гипотезу. Чем меньше р-уровень значимости, тем больше вероятность отклонить H0 гипотезу в пользу альтернативной Н1.

Если р-значение меньше 0.05, то можно смело отвергать Н0 гипотезу. Если р-значение больше 0.05, то считается, что у нас недостаточно оснований отвергать Н0 гипотезу.

**P-value** показывает вероятность отклонения выборочного среднего от среднего ГС при условии, что выборочное среднее рассчитано для выборки из ГС.

Вероятность некоторого события – это площадь под кривой распределения. Обычно отклонения бывают двусторонними, кроме редких односторонних исключений.

**P-значение** (англ. P-value) — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода).

Обычно P-значение равно вероятности того, что случайная величина с данным распределением (распределением тестовой статистики при нулевой гипотезе) примет значение, не меньшее, чем фактическое значение тестовой статистики.

**P-значение** – это наименьшее значение уровня значимости (т.е. вероятности отказа от справедливой гипотезы), для которого вычисленная проверочная статистика ведет к отказу от нулевой гипотезы.

p-значение меньше, тем лучше, поскольку при этом увеличивается «сила» отклонения нулевой гипотезы.

**Число степеней свободы** — это число категорий или переменных, которые вы анализируете в своем эксперименте.

Пример: в нашем эксперименте две категории результатов: одна категория для красных машин, и одна для синих машин. Поэтому в нашем эксперименте у нас 2-1 = 1 степень свободы. Если бы мы сравнивали красные, синие и зеленые машины, у нас было бы 2 степени свободы, и так далее.

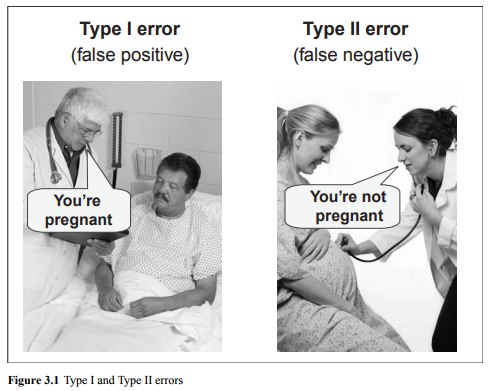
**Хи-квадрат** — это числовое значение, которое измеряет разницу между ожидаемыми и наблюдаемыми значениями эксперимента.

**Уровень значимости** показывает, насколько мы уверены в наших результатах. Уровни значимости записываются в виде десятичных дробей (таких как 0.01), что соответствует вероятности того, что экспериментальные результаты мы получили случайно (в данном случае вероятность этого 1%).

Ссылка на статью <http://datascientist.one/p-value/>

P-value не сообщает нам об уровне различий, он лишь позволяет сохранить или отвергнуть Н0 гипотезу.

**Ошибка первого рода** – это отклонение Н0 гипотезы, хотя она была верна. **Ошибка второго рода** – это не отклонение H0 гипотезы, хотя была верна Н1 (альтернативная) гипотеза.

Н0 – отсутствие беременности.