Дисперсия. Зачем она нужна?

Давайте посмотрим на два ряда чисел:

1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5 5 5 6 6 6 6 6 7 7 7 7 7 8 8 8 8 8 9 9 9 9 9
1 2 2 3 3 3 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 7 7 7 8 8 9

Что вы видите общего между этими двумя последовательностями? У них общий размах от 1 до 9, используются одни и те же числа и даже одинаковое среднее равное 5. Однако, если приглядеться к самим числам, то мы увидим, что в верхнем ряду одинаковое количество каждой цифры, т.е 5 единиц, 5 двоек и т.д. В этом случае мы можем говорить о равномерном распределении чисел. В случае же нижнего ряда, мы видим, что числа сгруппированы вокруг центра или среднего значения: больше всего пятерок, 4-ре четверки и шестерки, три тройки и семерки, и так далее. Т.е. в верхнем числовом ряду мы имеем ситуацию, *когда числа сильнее* ***разбросаны относительно своего среднего*** *арифметического*. А в нижнем же ряду разбросанность чисел не такая сильная, т.к у нас чаще встречаются “центральные” числа и реже крайние. Сравним это на графиках.

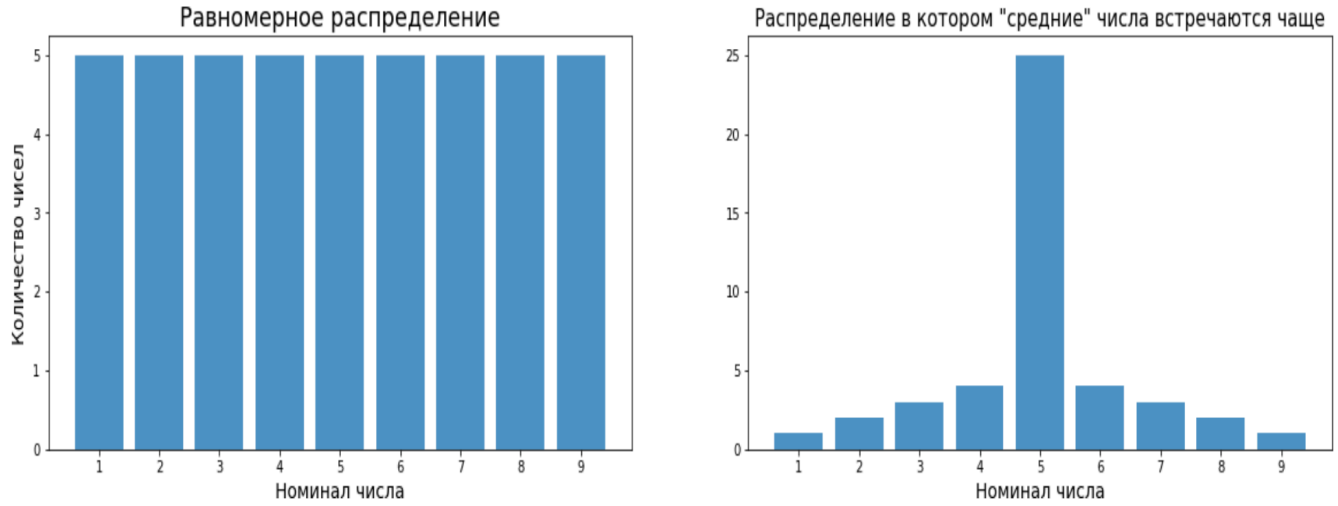


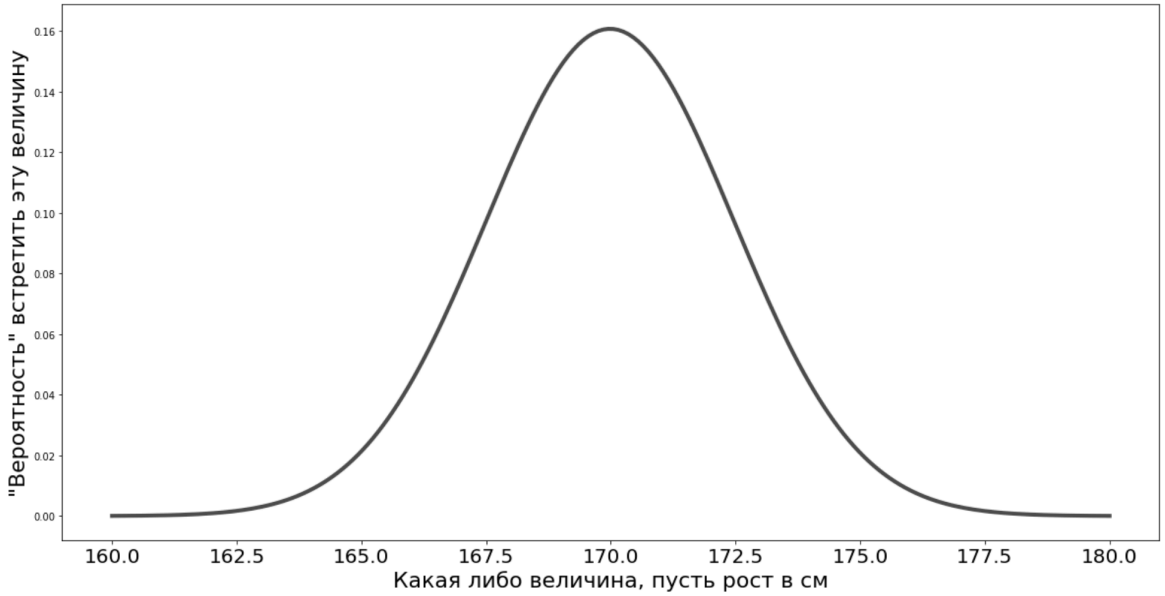
Рис.1 Числа слева сильно разбросаны. Числа на графике справа концентрируются вокруг среднего и в результате собраны кучнее.

Дисперсия выступает в качестве мерила *“разбросанности” чисел вокруг среднего*. Т.е чем дисперсия больше, тем больше будет чисел непохожих на среднее. Давайте посчитаем дисперсию в обоих случаях. Формула из Excel дает дисперсию 6,67 для первого ряда и 2.22 для нижнего, что согласуется с нашими ожиданиями, ведь в первом ряде числа разбросаны сильнее относительно среднего. Больше никакого смысла дисперсия не несет.

Если вы усвоили понятие дисперсии, то это уже хорошо, т.к больше дисперсию мы трогать не будем.

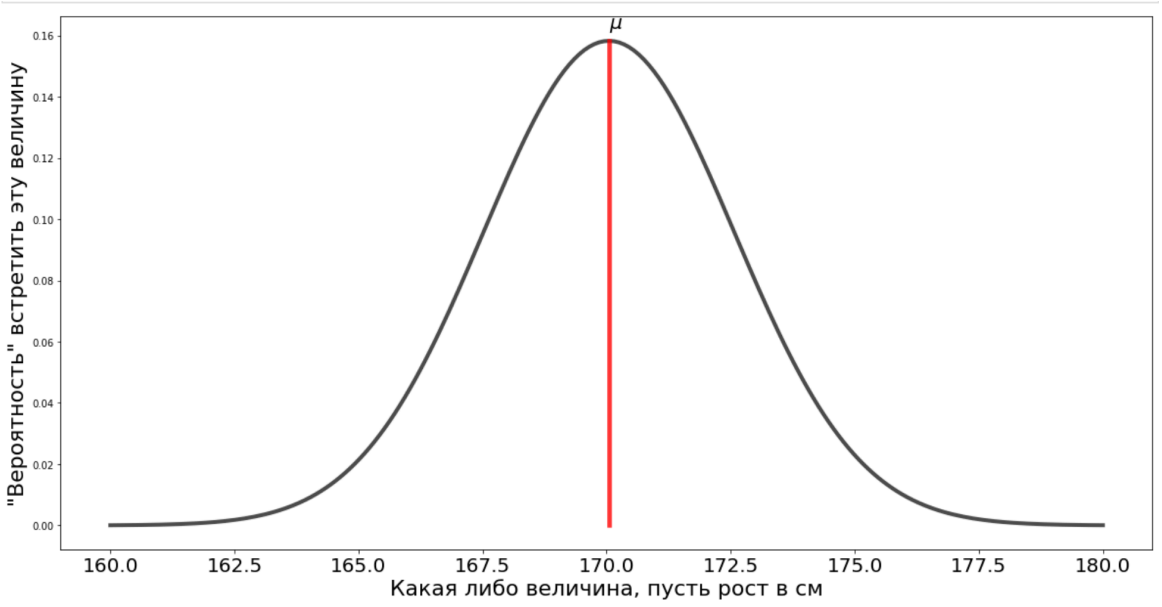
Стандартное отклонение или сигма

Что мы будем использовать дальше, так это корень из дисперсии, который называется стандартным отклонением. Почему мы так много крутимся вокруг двух этих чисел: **мат.ожидание (среднее)** и **сигма**? Потому, что зная только эти два числа уже можно строить такие графики (среднее=170 и сигма=2.5):



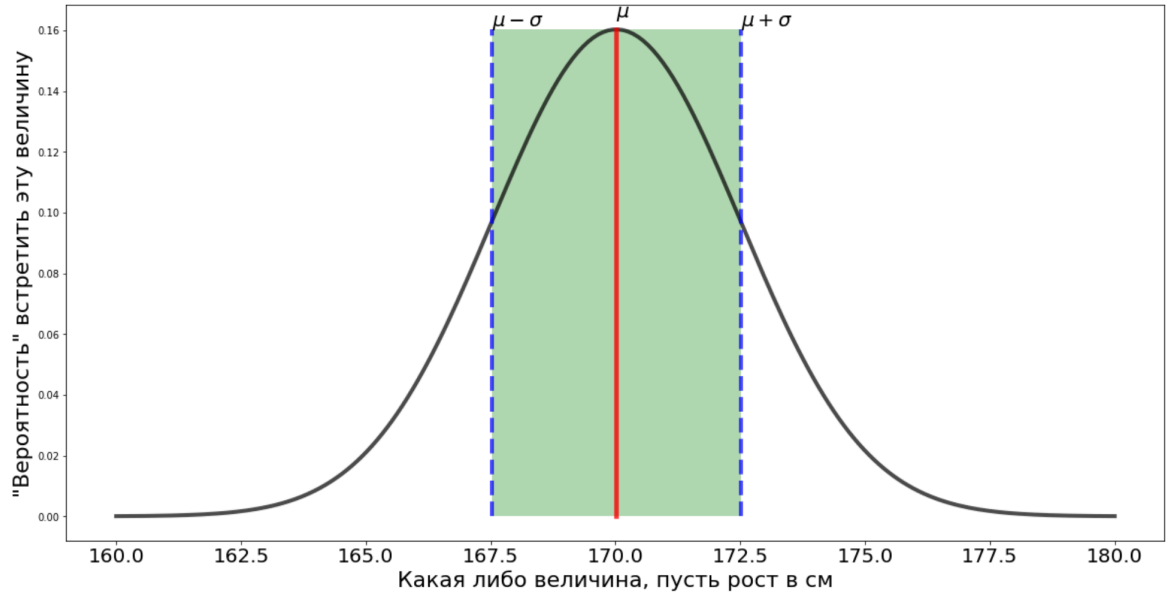
Кривая графика такого вида для нас ценна потому, что очень много случайных величин от психиатрии до машиностроения распределены именно по такому закону.

Идем дальше. Где же тут среднее, или мат.ожидание или **μ** (это всё одно и тоже)? А вот оно:

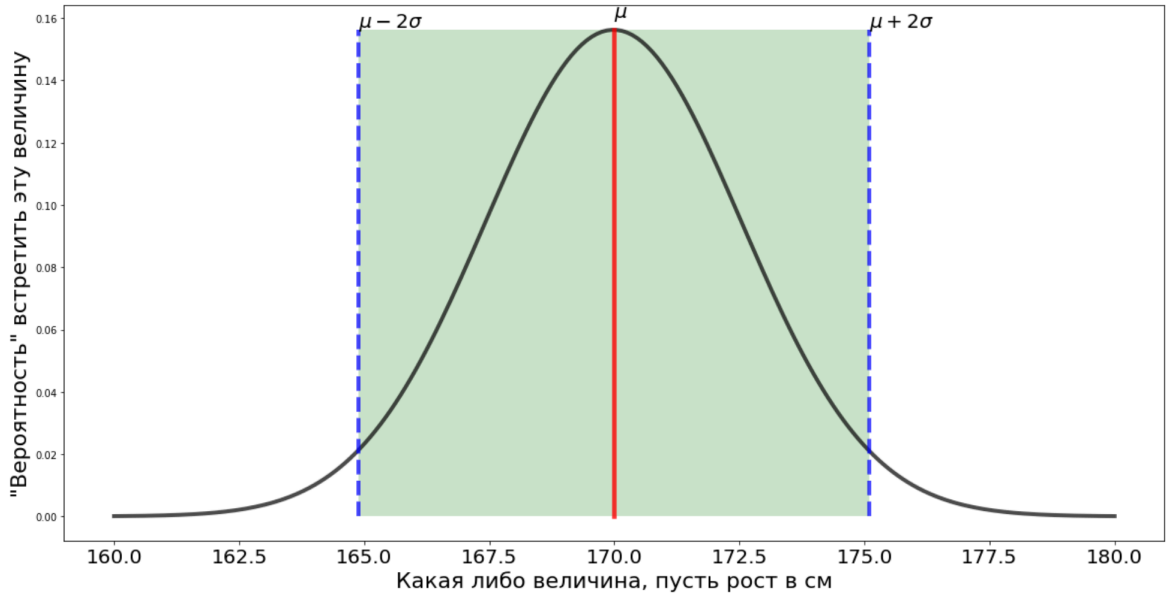


Теперь среднее арифметическое становится новой точкой отсчета. Мы можем от **μ** откладывать отрезки вправо и влево с каким-то шагом. Вот тут и пригождается

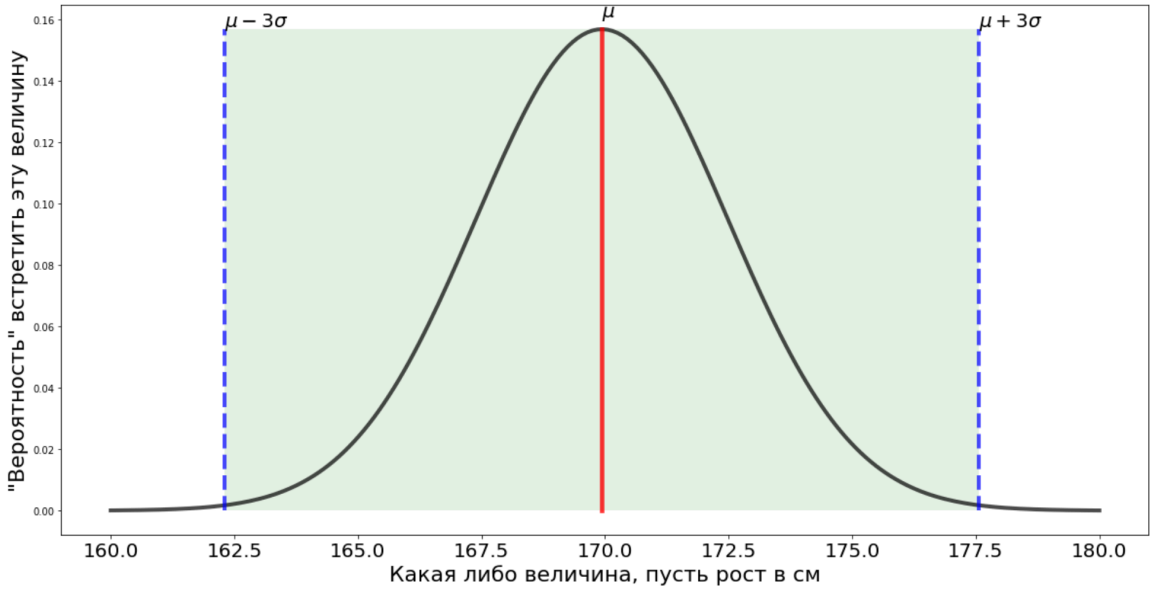
сигма. Оказывается если шагнуть влево на сигма (мы помним, что сигма=2.5) и вправо на сигма от красной линии (которая отвечает за среднее) то мы можем пометить вот такой диапазон:



В чем соль такого диапазона? Оказывается, что в такой диапазон попадает 68% измерений. Запомните это. Конечно, никто не ограничивает нас строить и другие диапазоны. Например в диапазон [**μ** - 2σ; **μ** + 2σ], т.е в нашем случае [170-2*2.5; 170+2*2.5] будет попадать 95% измерений:



Или например, можно строить 3-сигма диапазоны, т.е [**μ** - 3σ; **μ** + 3σ], т.е [170-3*2.5; 170+3*2.5], и сюда будет попадать уже 99% измерений. Зачастую 99% нас устраивают и мы на этом можем успокоиться.



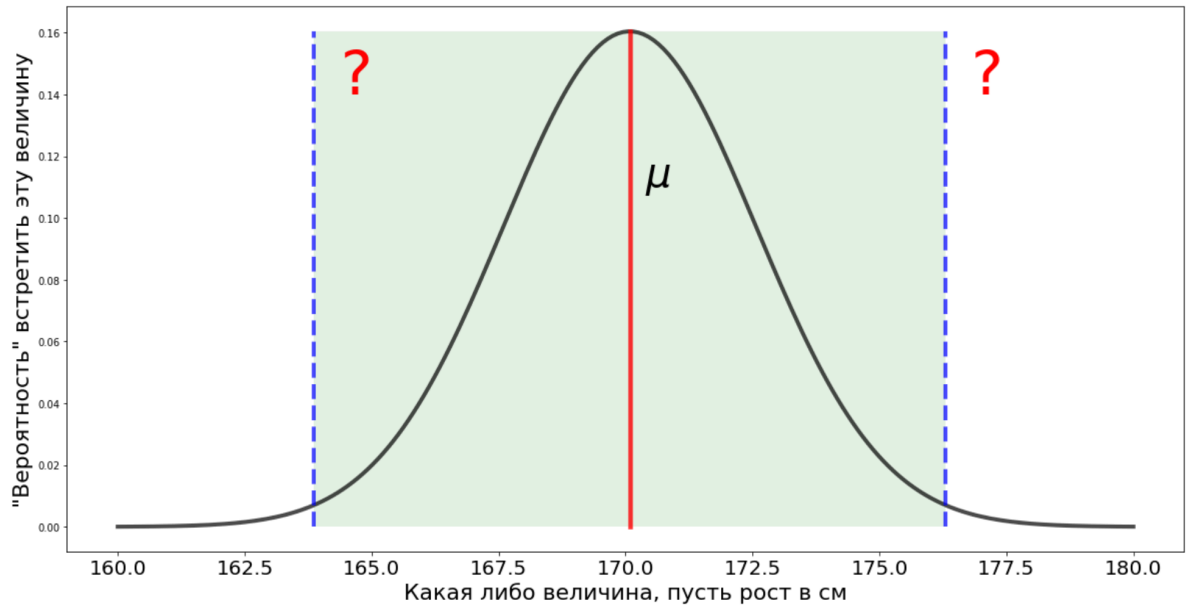
Связь с доверительными интервалами

Что мы сейчас проделали? Мы построили такие интервалы, которые позволяют находить, какой процент чисел помещается в заданных границах диапазона.

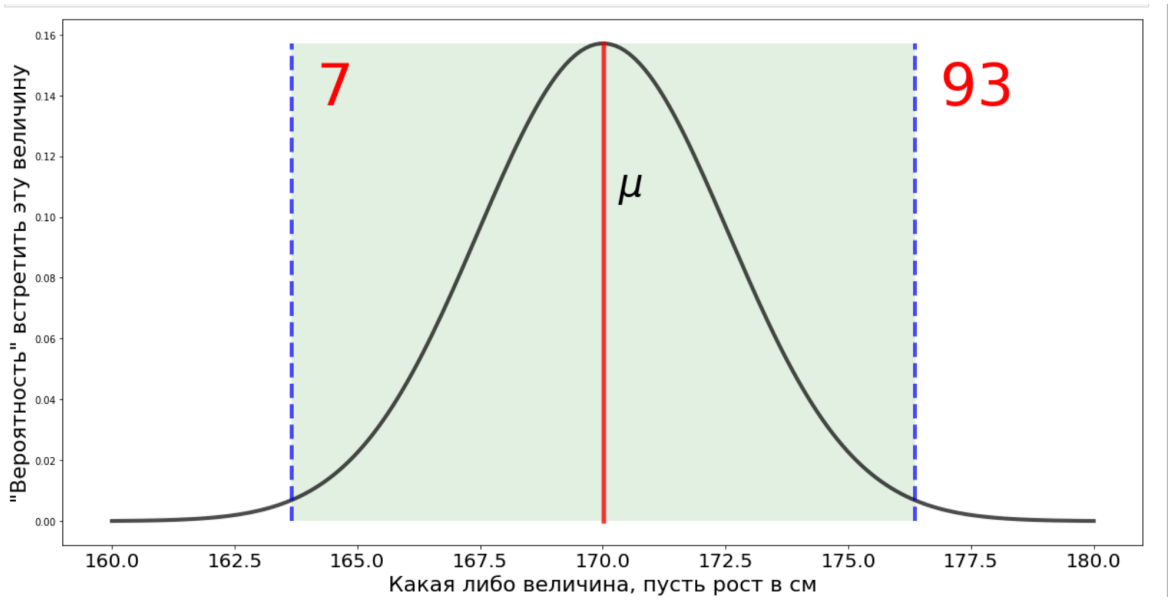
Сигма-интервал	На наших числах	Сколько процентов измерений включает такой диапазон
[**μ** - σ; **μ** + σ]	[170-2.5; 170+2.5]	68%
[**μ** - 2σ; **μ** + 2σ]	[170-2*2.5; 170+2*2.5]	95%
[**μ** - 3σ; **μ** + 3σ],	[170-3*2.5; 170+3*2.5]	99%

В принципе неплохо. Но мы не хотим на этом останавливаться. А что если хотим решать обратную задачу. **Т.е какой диапазон мы должны выбрать, чтобы данный диапазон включал 33%, 47% или 86%?** Т.е какие левая и правая границы[?; ?]. И как вообще мы должны их определять?

Давайте попробуем найти такой интервал для 86%.



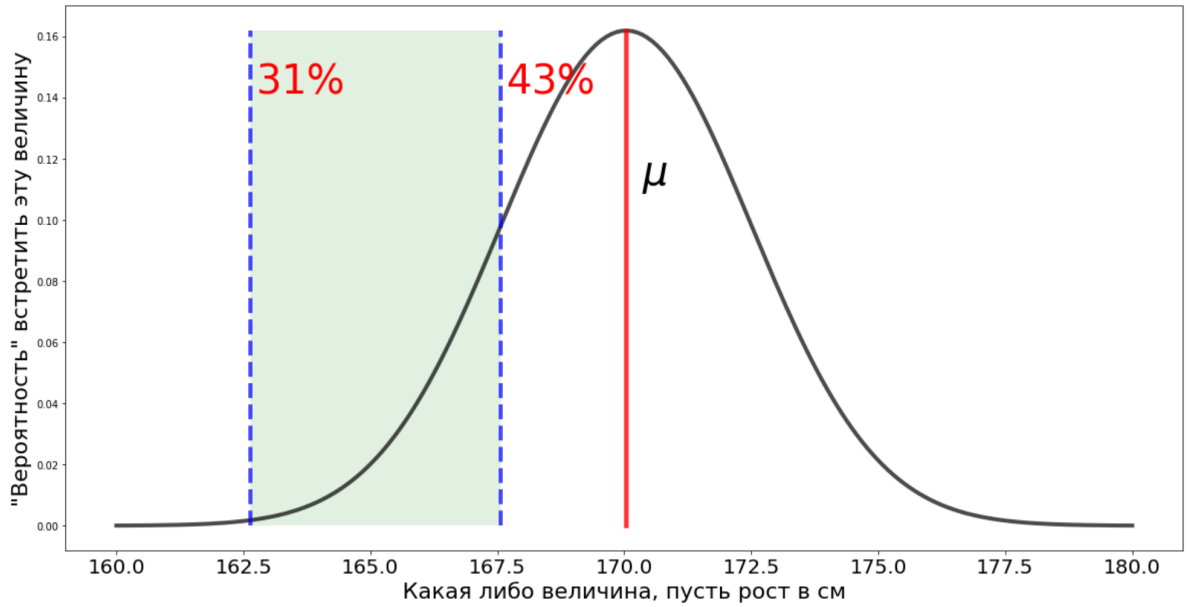
Давайте задумаемся еще об одном моменте. Сколько процентов измерений слева и сколько процентов измерений справа нам нужно отсечь, чтобы в итоговой зеленой зоне осталось 86%? Т.е мы хотим отсечь 14% (100%-86%) “ненужных” измерений. А т.к распределение симметричное, то нам надо отсечь 7% слева и 7% справа. 7% справа -- это 93%. Т.е между 7 и 93 помещается 86 процентов.



Excel приходит на помощь и предлагает помощь со встроенной функцией **NORM.INV** Как её использовать? Необходимо передать три аргумента: 1) процент, на котором нужно отсечь 2) среднее 3) сигма. Т.е NORM.INV(0.07, 170, 2.5) для левой границы и NORM.INV(0.93, 170, 2.5) для правой. Как итог, диапазон в который поместилось бы 86% измерений -- это [163.12; 176.18].

Таким образом мы можем строить диапазон, в который попадает любое интересующее нас количество измерений.

В статистике диапазон в (33%, 47%, 86 %, whatever) называется 1 - α диапазоном. **Самостоятельно** найдите, какие границы у такого диапазона. Ответ: [α/2, 1-α/2]
На самом деле можно и другие диапазоны строить при помощи NORM.INV, определяя верхнюю и нижнюю границы через [NORM.INV(0.31, 170, 2.5); NORM.INV(0.43, 170, 2.5)] :



Можно, но не надо.

Проверка гипотез

Давайте решим задание №1 из Д35.

1. Проверьте гипотезу о том, что мат. ожидание цены дома равно 1000

Лекционный материал даёт такую последовательность действий:

1. Формулируем задачу;
2. Формулируем нулевую и альтернативную гипотезу;
3. Задаем критическое значение;
4. Ищем нужный статистический критерий;
5. Ищем правильное распределение статистического критерия;
6. Проводим измерение;
7. Ищем p-значение;
8. Сравниваем его с критическим значением;

Пункт первый

*мат.ожидание в этой задаче это и есть среднее арифметическое цен домов (колонка A)
Задача уже сформулирована в условии, и нам надо понять, отличается ли мат.ожидание, которые мы посчитали от того, что есть в условии.

Пункт второй

Сформулируем нулевую и альтернативную гипотезу. Нулевая гипотеза должна отражать common sense, status quo или господствующее положение вещей. И мы собираем данные, чтобы опровергнуть это утверждение или наоборот доказать его. Условие задания задает status quo: мат.ожидание цены дома равно $1000. Альтернативную гипотезу можно сформулировать по-разному: мат.ожидание больше $1000, мат.ожидание меньше $1000 и мат.ожидание не равно $1000. Давайте выберем первое утверждение, и альтернативная гипотеза будет такая: на самом деле цена больше, чем $1000.

У нас есть колонка PRICE со 100 наблюдениями, проанализировав которые мы должны остаться либо с нулевой гипотезой, либо с альтернативной гипотезой.

Пункт третий

Задаем критическое значение. По-другому эту величину называют уровень значимости и обозначают буквой α. Его выбор остаётся на право того, кто проводит исследование, в разных задачах оно может быть разным. Обычно выбирают α=0.05, но иногда могут выбрать α=0.01 или 0.1. Мы выберем 0.05.

Пункт четвертый

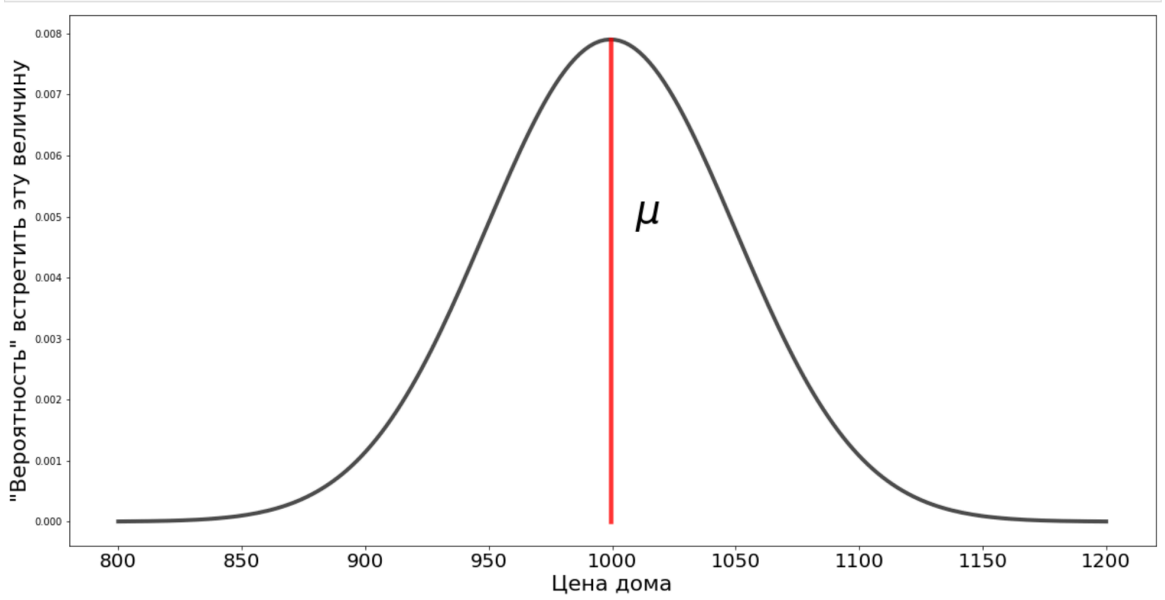
Ищем нужный статистический критерий, т.е по какой критерию или по какому фактору мы будем сравнивать. Ведь мы не можем просто сравнить всю колонку PRICE с $1000, ведь какие-то то цены будут больше, какие-то сильно больше, а какие-то меньше, чем $1000. В колонке Price присутствует весь плюрализм мнений. Поэтому нам нужна какая-то обобщающая характеристика. На самом деле этот пункт мы уже проделали, т.к само задание прямо на это указывает. Нужно проверять мат.ожидание всей колонки, т.е это одно число и как сильно оно отличается от цены указанной в нулевой гипотезе. Ключевое здесь слово: как сильно. Пока непонятно, как правильно определить слово “как сильно”, поэтому не пугайтесь.

Пункт пятый

Ищем правильное распределение статистического критерия.
Возможно, это вам покажется неудивительным, но статистический критерий (мат.ожидание, которое в нулевой гипотезе) тоже распределен по-нормальному распределению, как и многое в нашей жизни.
Теперь нам надо задать это распределение. Чтобы задать нормальное распределение нам надо знать два числа: мат.ожидание и сигму. Мат.ожидание у нас уже есть, по условию оно равно $1000. А какое же у него сигма? Это хороший вопрос, на него никто не знает ответа. В некоторых задачах, оно бывает известно по условию, но не в нашем случае.
Так как же быть с неизвестной сигмой? Статистика и тут приходит на помощь, и говорит, раз мы точно не знаем сигму, то давайте её оценим. А оценивать будем

следующим образом. Ведь у нас есть колонка с примерно 100 ценами, давайте найдем сигму по этим ценам. Нашли STDEV.S(A:A) = 404.38.
Вот тут мы должны выдохнуть и сказать, что **это сигма самих цен, но никак не мат.ожиданий цен**. Сигма для мат.ожиданий цен в $\sqrt{N}$ раз меньше, чем сигма самих цен, где N - это количество наблюдений, т.е N=66. Скорее всего, это непростой факт для понимания, поэтому дайте время ему настояться.

Итого мы имеем, что **μ** = 1000 σ = 404.38/$\sqrt{66}$ = 49.79
По этим двум числам можно построить распределение для статистического критерия:

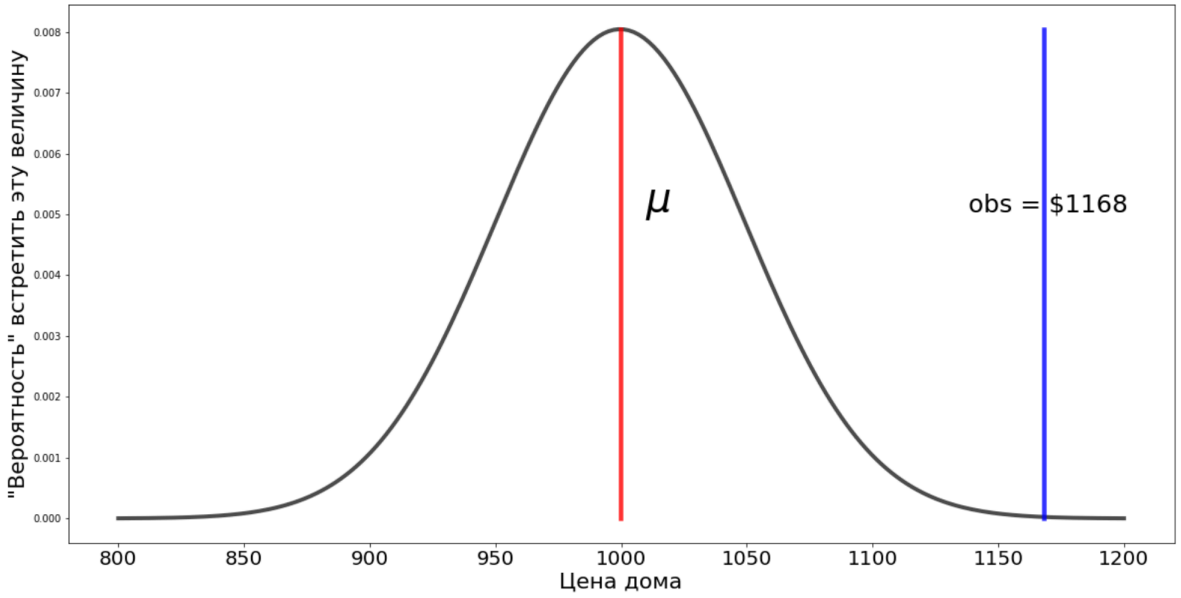


Обратите внимание, что график наверху называется **нулевым распределением**, т.к мы строили это распределение по параметрам нулевой гипотезы.

Пункт шестой

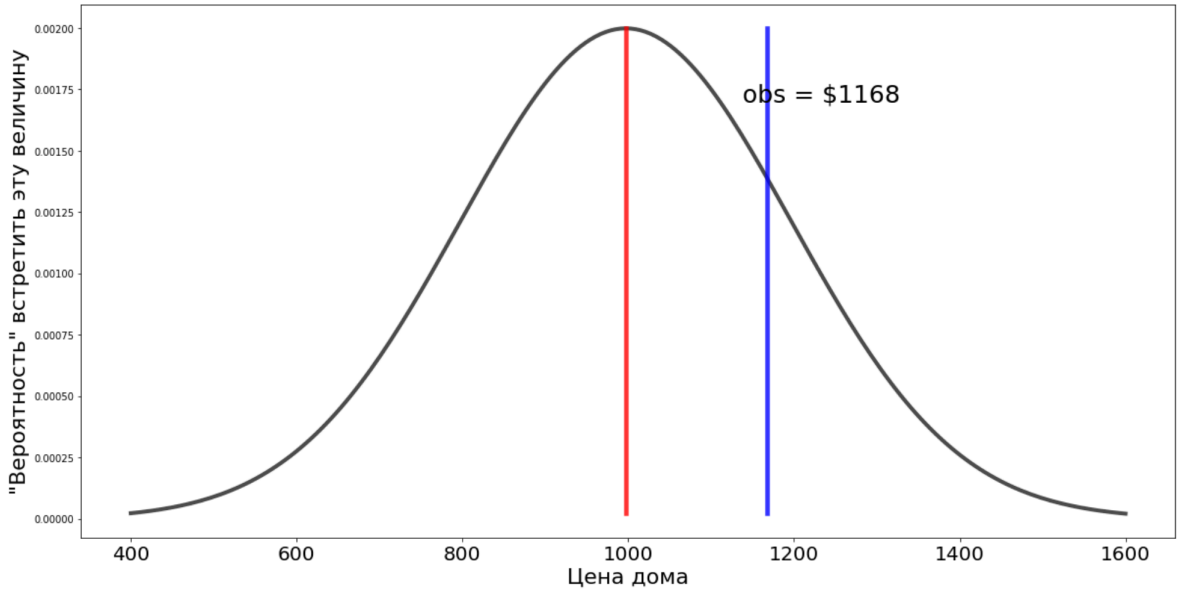
Проводим измерения.
Согласно данным в таблице по колонке PRICE, мат.ожидание набора чисел равно AVERAGE(A:A) = $1168. Что это число говорит о нулевой гипотезе, достаточно ли это число велико, чтобы опровергнуть нулевую гипотезу? Пока непонятно. Но мы интуитивно чувствуем, что если бы AVERAGE(A:A) = $50000, то скорее всего нулевая гипотеза точно не верна. Непонятно как поступать в случаях, когда AVERAGE(A:A) = $1200, $2000, $3000 или $900, $500?

Однако, давайте посмотрим, где наше измеренное значение лежит на кривой нулевого распределения.



Далековато, правда? Скорее всего, нулевая гипотеза неверна, т.к $1168 слишком далеко лежит от **μ**=1000. Но этого недостаточно, т.к мы хотим узнать насколько далеко наше измеренное значение удалилось от среднего. Давайте подумаем почему мы не можем использовать просто разницу 1168 и 1000, или насколько эта разница 1168-1000 отличается от 1000, т.е 1168-1000/1000?

Представим, что в наблюдений был бы заметен **больший разброс**, т.е сигма была больше. Обратите внимание, как изменились данные по оси OX. И эта разница 1168-1000/1000 или 1168-1000 останется такой же. Но мы видим, что 1168 уже лежит ближе к среднему, и скорее всего не стоит отвергать нулевую гипотезу.



Держите в уме тот факт, чтобы понять насколько наше измерение лежит далеко от среднего мы должны учитывать сигму.

Пункт седьмой

Так как же понять насколько 1168 лежит далеко от 1000? Вот здесь и приходит на помощь p-value. По-другому еще называется достигаемый уровень значимости. Еще это число говорит о вероятности того, насколько мы могли бы получить наше наблюдение случайным образом.
p-value в Excel считается по NORM.DIST(s, **μ**, σ), где s -- это то, что мы намеряли, т.е $1168, **μ** -- мат.ожидание в нулевой гипотезе ($1000), σ -- сигма в нулевой гипотезе (49.79).
T.e pvalue=NORM.DIST(1168, 1000, 49.79, False) = 0.0000270

***Спасибо внимательным студентам. На самом деле, перед тем как считать p-value необходимо стандартизировать данные.**

pvalue=NORM.DIST((1168-**μ**) / $\sqrt{\sigma}$, 0, 1, False)

Пункт восьмой

Мы получили pvalue=0.0000270. Оно очень маленькое и говорит о том, что вероятность получить измерение $1168 случайным образом было мало.
Чтобы сделать заключительный шаг, надо сравнить с pvalue с α. Если pvalue < α, то мы говорим, что мы отвергаем нулевую гипотезу на уровне значимости α. А если pvalue >= α, то мы говорим, что нулевая гипотеза верна либо данных недостаточно, чтобы опровергнуть нулевую гипотезу.

Как итог, мы отвергаем нулевую гипотезу, что мат.ожидание цены дома равно $1000.

Такие вот дела.