Автор выражает благодарность Новиковой Евгении за подаренную идею

Распределения Гаусса (оно же нормальное) и Стьюдента – два очень похожих распределения. Оба колокообразные, симметричные. Когда нужно применять одно, а когда другое? А когда вообще Пуассона? Обо всём этом мы поговорим в этой методичке.

В Подмосковье есть прекрасный город Зарайск, в которым живёт один человек, допустим, Оля. Страдая от бессонницы, она каждую ночь считала цветы на клумбе у себя под окном. Каждую ночь она записывала свои тетради в себе блокнот. Так продолжалось k ночей.
Число цветков на клумбе одно и то же (будем считают, что они не расцветают и не вянут). А вот Оля считает с некоторой погрешностью: на улице темно всё же.

Вопрос: какая погрешность Олиных измерений? И какое там распределение числа людей? Гаусс? Стьюдент? Пуассон?

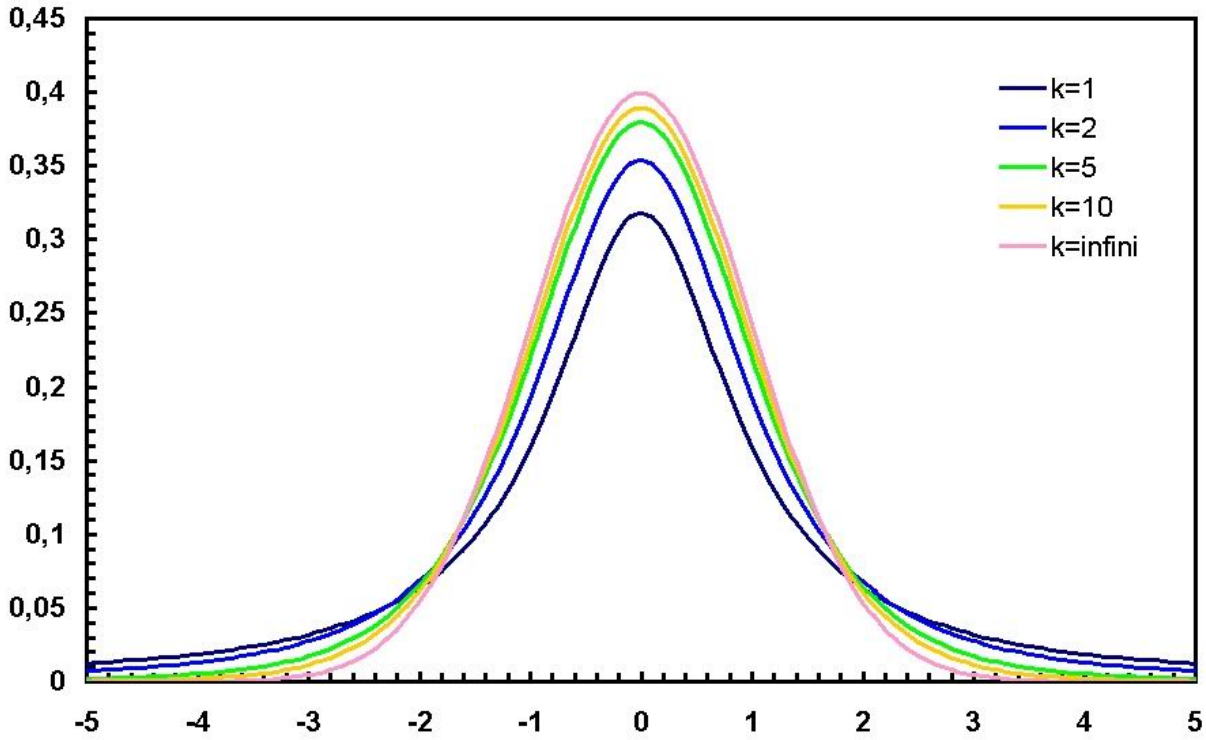
Спойлер: ни то, ни то, ни то.

Гаусс выглядит на первый взгляд правдоподобным: простая колокообразная функция, среднее у неё будет, конечно, в тройке. Только мы не знаем дисперсию σ^2, которая нам нужна для построения Гаусса. Думается, что σ^2 зависит от наблюдателя, который может считать неточно: если у нас наблюдает дед со зрением -10 без очков, то дисперсия у него будет велика, если у нас молодая Оля в очках, то дисперсия будет мала. Только мы её всё равно не знаем…

И это незнание дисперсии выносит приговор Гауссу.
Если мы проводим измерение (да хоть одно, не обязательно серию) какой-то величины с помощью измерительного прибора ***с известной дисперсией (!!!)*** - то плотность вероятности будет нормальное распределение.
Скажем, пусть у нас есть кирпич, и мы хотим измерить его длину линейкой. Но так как у нас руки кривые, мы получим результат не абсолютно точный. И пусть нам Всевышний сообщил, что погрешность (она же корень из дисперсии) будет σ (а дисперсия - σ^2). Тогда функция плотности вероятности будет Гаусс!

А вот если дисперсия нам неизвестна (как в данном случае), то будет Стьюдент.

Напомним, что это за распределение такое. Для начала – как выглядит график функции плотности вероятности? Похож на Гаусса, но более широкий.



Тем больше k, тем колокол уже, и при k->+∞ Стьюдент стремится к Гауссу.

Распределение Стьюдента играет важную роль в статистическом анализе и используется, например, в *t*-критерии Стьюдента для оценки статистической значимости разности двух выборочных средних, при построении доверительного интервала для математического ожидания нормальной совокупности при неизвестной дисперсии, а также в линейном регрессионном анализе. Распределение Стьюдента также появляется в байесовском анализе данных, распределённых по нормальному закону.

Отметим, что если число ночей k достаточно велико, то Оля может посчитать дисперсию экспериментально: как средний квадрат отклонения от среднего. А зная дисперсию, можно и о Гауссе вспомнить, для которого нам как раз нужно знать дисперсия.
И Стьюдент также при больших k стремится к Гауссу – раз дисперсия экспериментально почти известна (всё-таки не абсолютно точно), то и распределение будет почти Гауссовым.
А вот если k будет мало – то Стьюдентов колокол будет широким. Хотим больше точность – делайте больше измерений. Вспоминаем Митина ☺

Итак, в споре Гаусс vs Стьюдент разобрались. Небольшая ремарка: и Гаусс, и Стьюдент – непрерывные распределения, а число цветков на клумбе дискретно.

Что же, надо будет Стьюдента сделать дискретным, построив в виде столбчатой диаграммы ☺

А вот распределение Пуассона само по себе имеет дискретную природу, и кто-то наверняка скажет, в исходной задаче ни Гаусс, ни Стьюдент, а Пуассон. Что же, он будет прав, если должным образом уточнить условие.

Итак, меняем условие – теперь не цветы, а люди.
Представим, что 30 тысяч жителей Зарайска (кроме самой Оли) каждую ночь играют в лотерею: погулять им ночью под Олиными окнами или нет. Каждый гуляет с вероятностью 1/10000, вероятность выигрыша каждого жителя не зависит от выигрышей других – все независимы, все тупо генерят число от 1 до 10000 рандомайзером, если 1 - гуляют. Тогда в среднем как раз будет под Олиными окнами три человека.
Это первое отличие. Второе: Оля фиксирует всех абсолютно точно, никакой погрешности, связанной с измерениями, нет.

Вот в этом случае будет чистый Пуассон. В частности, мы уже можем сделать вывод, что дисперсия Олиных записей будет 3 человека^2, а погрешность - $\sqrt{3}$ человек.

Давайте обсудим разницу между Гауссом-Стьюдентом и Пуассоном.

В Гауссе-Стьюденте у нас была одна истинная величина, которую мы измеряли кучу раз с некоей случайной погрешностью измерения. Ну вот у нас, например, кирпич (в форме идеального параллелепипеда), длину которого мы меряем линейкой. У кирпича есть одна истинная длина, но мы линейку не под прямым углом каждый раз ставим и всё равно есть некая погрешность измерений. Это Стьюдент!
(Гаусс – то же самое, но если нам известна погрешность).

А Пуассон про другое – нет истинного значения, оно меняется каждый раз от ночи к ночи, но зато мы меряем всё точно, без погрешностей.

А что будет, если и то, и то – Оля считает число людей, которое каждую ночь разное, да ещё сама вносит погрешность измерения?
В этом случае у нас два источника рандома. Будет ухудшенная (более широкая) версия Стьюдента (или ухудшенная версия Пуассона). Предоставить точную формулу функции плотности вероятности увы, не могу.

Ещё пример. Помните, нас на 14 ядерном праке (это где калиевое удобрение, Гейгер и 200 измерений) просили проверить, что распределение Пуассоново? Оно действительно будет Пуассоново, если детектор измеряет число частиц абсолютно точно. Но если он, например, не цифровой, а виде стрелки на циферблате, где ещё вдобавок существует погрешность считывания – уже будет не совсем Пуассон, а ухудшенный (более широкий) Пуассон. Если погрешность будет большой – то она может даже превзойти погрешность, связанную с рандомом внутри калиевого удобрения (ту, которая корень из матожидания) – и тогда будет уже совсем не Пуассон, а нечто отвратное.

Давайте ещё раз, чтобы было точно понятно.
Представим, что у нас два счётчика Гейгера – один со средней погрешностью 1 частица, другой с погрешностью 5 частиц.

Число частиц в удобрении, распавшихся за 30 сек	Что намерил первый Гейгер	Что намерил второй Гейгер
20	20	15
27	28	30
25	24	30
18	19	14
30	30	35
24	25	27
19	19	15
22	23	19
26	27	25
20	19	22
17	17	21
25	24	30

Давайте анализировать. В первом столбце у нас ***абсолютно точно*** распределение Пуассона – это прямо следует из теории радиоактивного распада. Во втором столбце у нас немного ухудшенный Пуассон – распределение будет похоже на Пуассона, но не в точности оно **– *из-за погрешности детектора!*** Ну а во втором будет полный разколбас из-за плохого детектора.

Поэтому, когда на 14-том праке говорят «проверьте, что распределение Пуассона», это не для того, чтобы проверить теорию радиоактивного распада (она и так верна), а для того, чтобы убедиться, что детектор работает корректно и не вносит заметной погрешности.