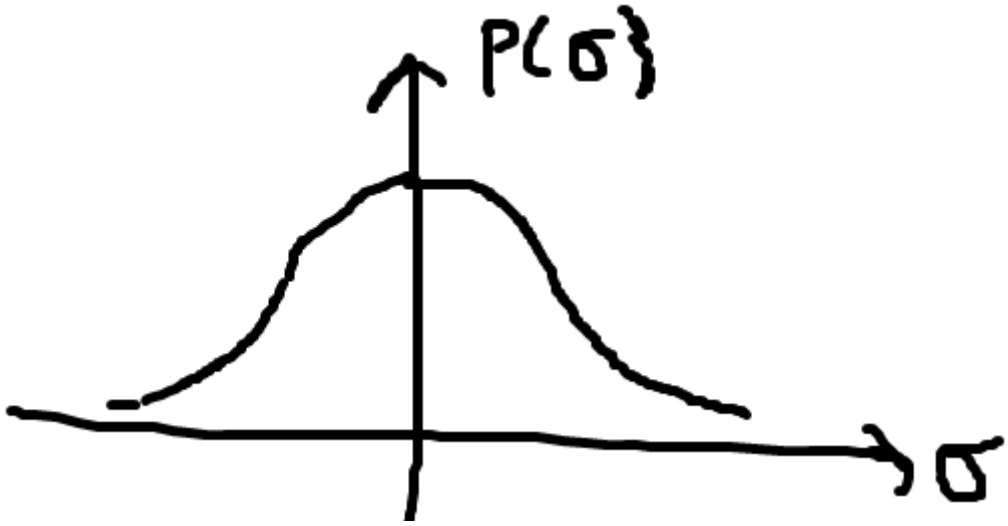
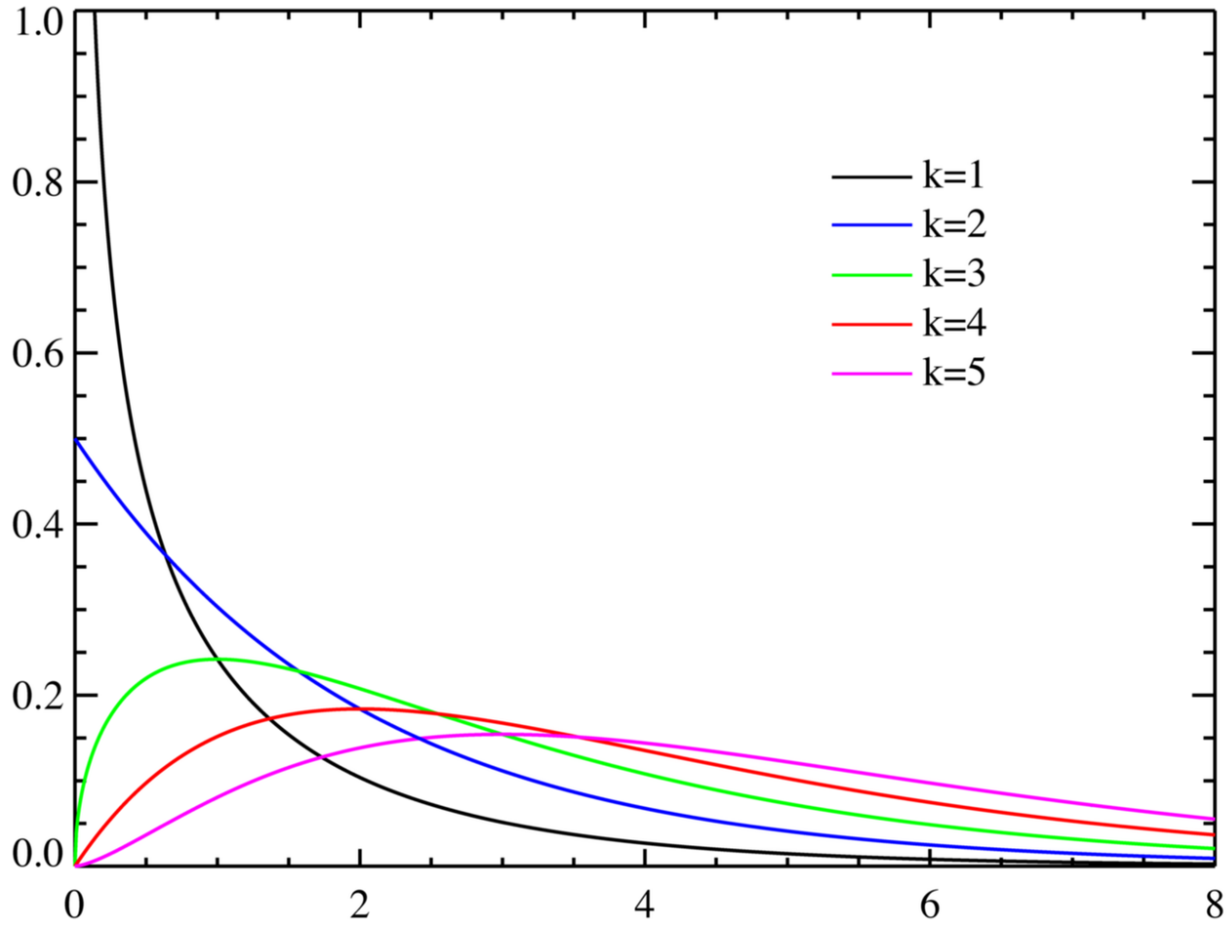
Хи-квадрат распределение.

Представим себе, что мы экспериментально измеряем некую величину. Мы предполагаем, что погрешность (разность экспериментального и истинного значений) подчиняется нормальному распределению с σ=1.

Т.е. наибольшая вероятность получить погрешность 0. Чем больше по модулю погрешность, тем меньше вероятность её получить.



Это был график плотности вероятности погрешности, а теперь найдём плотность вероятности ***квадрата*** погрешности, т.е. дисперсии. Это будет чёрный график (при k=1):



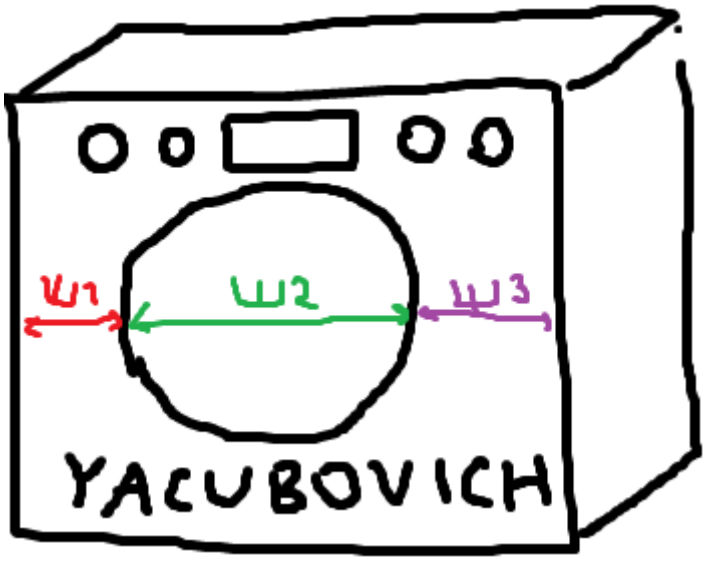
Как мы видим, вновь максимум в нуле. Наиболее вероятная дисперсия – 0. Полученный график – это хи-квадрат распределение при значении параметра ν=1.

Теперь представим, что мы измеряем величину Ш, которая равна сумме Ш1+Ш2+…+Шν. Измеряем мы, измеряя каждую Ш-ку, а потом суммируя. Экспериментально измеряя каждую Ш-ку, у нас получается некая погрешность. Нам может повезти и мы очень точно измерим Ш1, а вот с Ш2 погрешность будет достаточно большой. Интересно, на сколько мы ошибёмся в итоге?
Вот это и есть хи-квадрат распределение с заданным параметром ν. Давайте посмотрим на наши графики. При ν=2 у нас по-прежнему максимум в нуле. Но уже при ν=3 он становится чем-то около 1. Что это означает? Что, скорее всего, дисперсия полученного нами Ш будет что-то около 1. Мы немного накосячим при измерении Ш1, немного при измерении Ш2, немного при измерении Ш3, и в Ш все эти косяки сложатся, и в итоге окажется, что вероятность практически точно измерить Ш очень мала, а ошибиться на единичку – вполне себе велика.
По мере роста ν у нас всё больше возможностей для косяков. Именно поэтому пик хи-квадрат-распределения постоянно сдвигается вправо.

Формула для хи-квадрат распределения очень сложная:

$$p_{\chi^2_\nu} = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}$$

Пример для понимания. Покупатель покупает стиральную машину в магазине. В руках у него линейка, которой он последовательно обмеряет ширину стиралки (линейка короткая, так что ему пришлось маркером сделать две отметки на стиралке, разбив ширину стиралки на три отрезка, которые он и померил).



Все измерения покупателя подчиняются нормальному распределению с дисперсией 1 $см^2$.
Мы – консультант магазина. Мы абсолютно точно знаем, что ширина стиралки – 51 см, а ширина прохода в ванную покупателя – 53 см. Какова вероятность, что покупатель после своих измерений не купит стиралку по причине кажущегося ему невлезания в проход?
Ответ: чтобы он не купил, нужно, чтобы он ошибся более, чем на 2 см (а дисперсия тогда должна быть больше 4 $см^2$). Надо подсчитать интеграл от 4 до бесконечности от хи-квадрат-распределения с $\nu=3$. т.к. покупатель делает 3 измерения. (Т.к. все величины измеряются или в см, или в $см^2$, размерность можем выкинуть).
(Пример не совсем точный, потому что не учитывает, что погрешность может быть с плюсом, а может с минусом. Чтобы сделать пример точным, будем считать, что покупатель, намеряя меньше 49 см, также откажется от стиралки).
Соответственно, вероятность покупки – интеграл от того же хи-квадрат распределения с $\nu=3$, но уже от 0 до 4.

Второй пример посложнее. Та же ситуация, но пришёл другой покупатель, у которого руки трясутся меньше. Дисперсия нормального распределения его измерений будет уже 0,25 $см^2$ (т.е. $\sigma=0,5$ см). Как нам теперь найти вероятность НЕпокупки?
Ответ: перейти к новым единицам измерения, полусантиметрам. Тогда $\sigma=1$ вновь будет единичной. Нам вновь надо будет интегрировать хи-квадрат-распределение при $\nu=3$ (число измерений осталось то же, 3), но вот пределы интегрирования будут не от 4 до бесконечности, а от 16 от бесконечности. Почему 16? Покупатель должен ошибиться уже на 2 сантиметра = 4 полусантиметра, а дисперсия будет квадрат погрешности, т.е. 16.

Заметим, что интеграл от 16 до бесконечности, конечно, будет меньше, чем интеграл от 4 до бесконечности. Что вполне понятно: у второго покупателя более точные измерения и больше вероятность получить истинный результат «стиралку надо брать» (мы-то знаем, что она пролезет…)

Распределение Стьюдента
Оно нужно для доверительных интервалов.
В прошлый раз мы были в роли консультанта, знали заранее ширину стиральной машины и угорали над покупателем, который пытался там что-то померить. Распределение Стьюдента о более реалистичной задаче: мы не знаем истинную ширину стиралки, потому что мы теперь в роли покупателя. Мы предусмотрительно захватили с собой длинную линейку (рулетку, если вам она более нравится) и провели целых 5 измерений одной и той же физической величины – ширины стиралки (в хи-квадрат мы один раз мерили 3 разных отрезка, здесь 5 раз один):
51,8 см

51,7 см
52,5 см
53,1 см
50,4 см
Хм. Среднее 51,9 см. Должна пролезть. Но мы один раз намеряли более 53,1 см. А вдруг не пролезет? Нужен доверительный интервал. Если, скажем, 98%, что пролезет, мы стиралку купим.

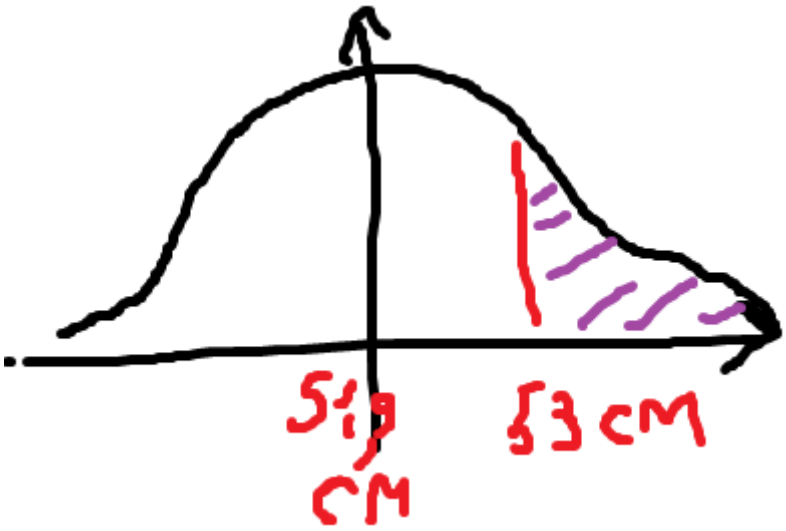
Та же проблема у физика-ядерщика. Он измеряет массу какой-то элементарной частицы. У него вышло
938,17 МэВ
938,57 МэВ
938,97 МэВ
938,87 МэВ
938,77 МэВ
В среднем 938,67 МэВ.
Он хочет понять, открыл ли он новую частицу или это просто протон, у которого масса 938,27 МэВ. Также нужен доверительный интервал.

Для этого и нужно распределение Стьюдента. Его график похож на нормальное распределение.
Формула:

$$p_{T_\nu} = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

Как и в случае с хи-квадрат-распределением, мы считаем, что все измерения были проделаны с дисперсией 1, иначе нужно всё измерять в таких единицах измерения, чтобы дисперсия была одна единица измерения (как во втором примере с хи-квадратом).
Не будем думать, откуда она взялась, а поймём, как ей пользоваться.

Покупатель, зная эту формулу, чертит распределение Стьюдента с ν=5, за его центр принимая среднее арифметическое того, что он намерил – 51, 9 см:



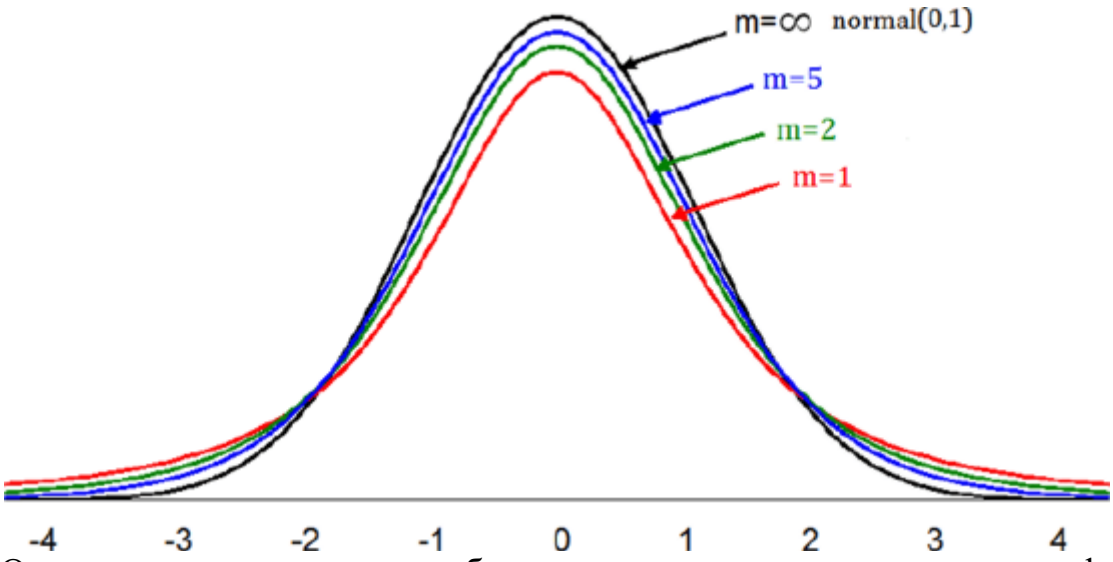
Тогда площадь фиолетовой области (которую мы, покупатель, вычислим обыкновенным интегралом) нам и укажет вероятность того, что стиралка в проход не влезет (на основании наших наблюдений).

Аналогично делает ядерщик. Строит своё распределение Стьюдента с ν=5, сдвинутое так, чтобы центр был на 938,67 МэВ:

Ну а далее двигает две вертикальных отрезка слева и справа в зависимости от того, какой доверительный интервал ему нужен. Площадь зелёной фигуры должна быть α, площадь жёлтой – β, площадь белой – γ. Это вы уже видели в главе 2 ☺

Да, распределение Стьюдента как раз и есть центральная статистика. Думаю, что в главе 2 эти понятия различают, потому что не всегда центральной статистикой будет распределение Стьюдента, но если не оговорено иное – юзайте его.

Чем больше ν, тем уже будет распределение Стьюдента и тем явнее будет выражен пик. Это вполне естественно: чем больше мы проделали измерений одной и той же величины, тем больше вероятность, что истина будет где-то поблизости от среднего арифметического.



Однако отметим, что при ν->бесконечности мы не получим дельта-функции, как нам бы хотелось (тогда мы бы 100% утверждать, что истина есть среднее арифматическое), а получим лишь нормально распределение.
Тем самым распределение Стьюдента представляет собой как бы ухудшенную версию нормального распределения, более широкую версию нормального распределения. Для нас это не есть хорошо, потому чем уже распределение Стьюдента, тем точнее наша оценка в виде среднего арифметического. Выход один – повышать ν, но даже в этом случае нашим «потолком» будет нормальное распределение.

Распределение Пуассона.
Представим себе, что мы n раз участвуем в лотерее, шансы выиграть в которой p. p<<1 (как это обычно бывает), но мы не растерялись и играем в неё очень много раз (n>>1). Устремим p к 0, а n к бесконечности так, чтобы их произведение осталось постоянным – θ. θ, как можно понять, маотжидание числа наших выигрышей – за одну попытку матожидание будет p, а за n попыток np=θ.

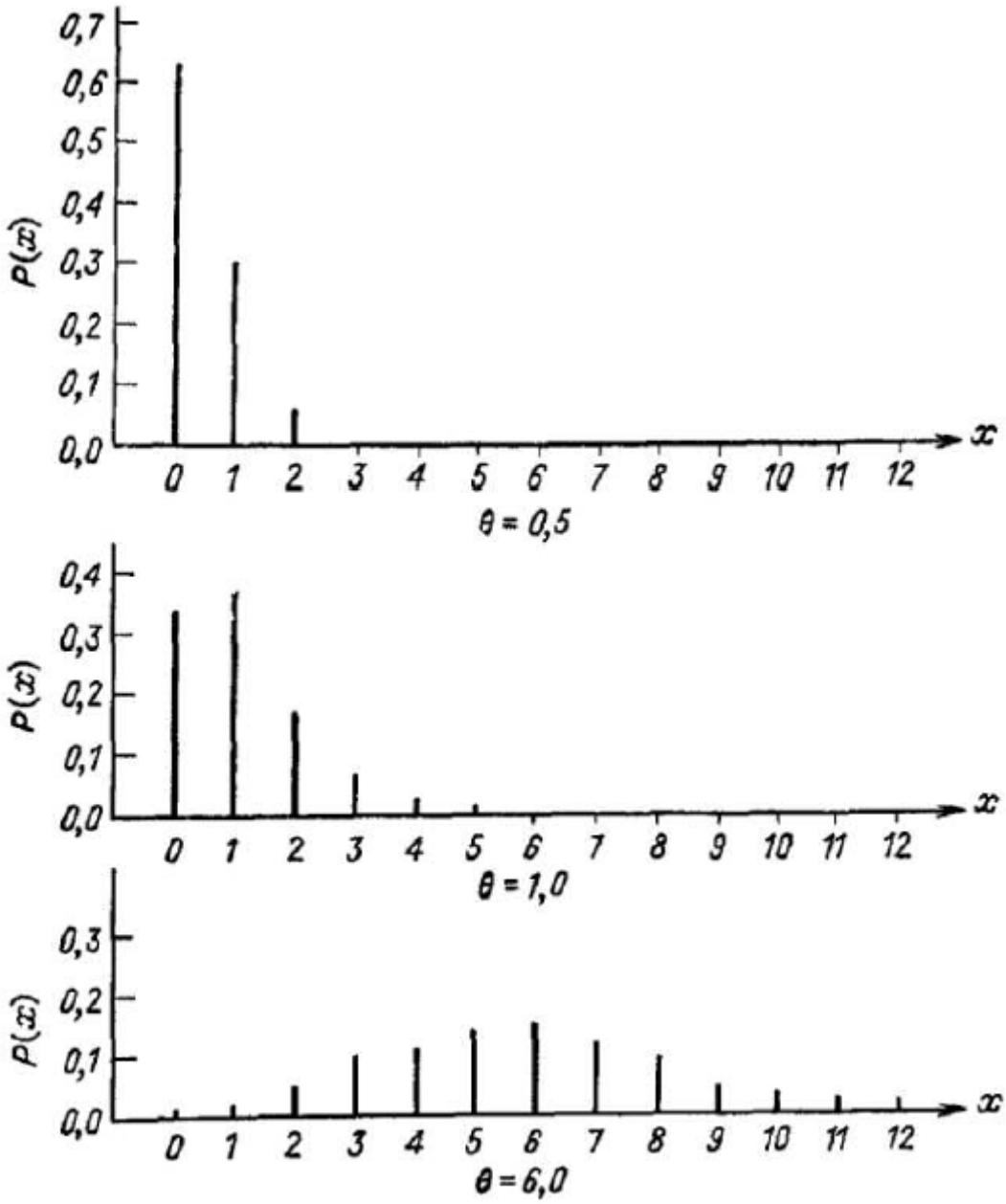


Рис. 1.2: Распределения Пуассона для $\theta = 0.5, 1.0, 6.0$.

Отметим, что распределение Пуассона – дискретное распределение, оно определено только при целых значениях аргумента.

Видно, что при θ=0,5 мы, скорее всего, ни разу не выиграем. При θ=1, скорее всего, мы победим один раз, но вероятность не победить ни разу тоже не мала и составляет около 1/3. Да, если вы 1000 раз сыграете в лотерею, где шанс выиграть 1/1000, вероятность уйти ни с чем 1/3. Грустно, но что ж поделать.

Ну а при θ=6, мы, скорее всего, выиграем 5 или 6 раз.

То, что матожидание будет θ, очевидно (см. выше). Прекрасным неочевидным свойством распределения Пуассона будет то, что тете (да не тёте, а тете. Тете. Греческой букве ☺) будет ещё равна и дисперсия:

$$M(X) = \theta, \quad D(X) = \theta.$$

Форма распределений сильно меняется с изменением параметра. При малых θ распределение является существенно асимметричным, но уже по достижении $\theta \approx 10$ вероятности симметризуются. На рис.1.3 показан случай $\theta = 9$. Хорошо видно, что функция плотности вероятности распределения Гаусса с теми же, что у распределения Пуассона, матожиданием и дисперсией вполне может служить огибающей вероятностей Пуассона. При дальнейшем увеличении θ формы двух распределений продолжают сближение.

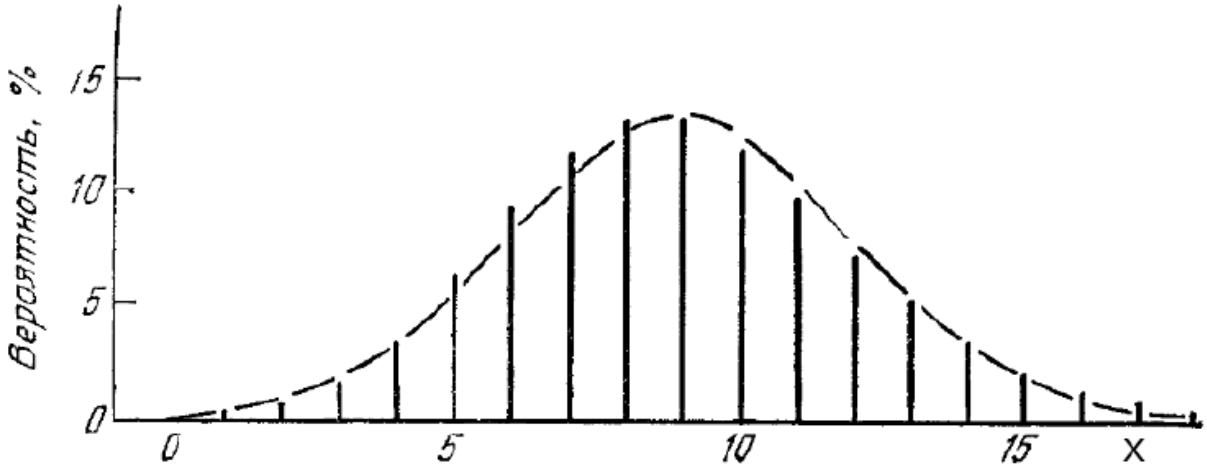


Рис. 1.3: Распределения Пуассона для $\theta = 9.0$. В качестве огибающей нарисована плотность вероятности Гауссова распределения с такими же матожиданием и дисперией.

Напоследок формула для расчёта распределения Пуассона:

$$P(x) = \frac{e^{-\theta}\,\theta^x}{x!} \quad \text{при } x = 0, 1, 2, \ldots .$$

Несмотря на то, что функция от x определена для любых неотрицательных x, смысл для распределения Пуассона она имеет только для целых x, потому что, напомню, распределение Пуассона – это дискретное распределение! Мы не можем выиграть в лотерею полраза или полтора, можем только целое число раз.

Подведём итоги:
Хи-квадрат: мы знаем истинное значение (ширины стиралки), это распределение ошибки покупателя. Чем больше ν (число составных отрезков), тем ***больше*** вероятность у покупателя накосячить.
Стьюдент: мы не знаем истинное значение, мы знаем выборку, распределение Стьюдента – это распределение плотности вероятности истинного значения по нашей выборке. Чем больше ν (число измерений одной и той же величины!), тем ***меньше*** вероятность у покупателя накосячить.
Пуассон: дискретное распределение про лотереи, характеризует наши успехи при многократной игре в очень маловероятную лотерею.