По критерию Пирсона существует 100500 статей и обучающих видео. «Заботать эту тему не составит труда» - подумал я в своё время и жестоко ошибался. 95% всего контента сводится к «иди и делай так». А почему – не объясняется. Лучшая ли это методичка по Пирсону, что есть? Не знаю. Но точно самая оригинальная. Обсудим те моменты, которые не проговариваются другими.

Для начала заметим, что тестов Пирсона, они же хи-квадрат несколько:

1. Тест на гомогенность (test of homogeneity, он же goodness of fit) — непараметрический, одновыборочный тест, который проверяет соответствие наблюдаемого распределения категориальной случайной величины некоторому эталонному распределению. В Python реализован функцией `scipy.stats.chisquare`.
2. Тест на независимость (он же test of independence/association) — непараметрический, одновыборочный тест, который проверяет наличие связи между двумя категориальными переменными. В Python реализован функцией `scipy.stats.chi2_contingency`.
3. Тест для дисперсии — параметрический (параметр — дисперсия), одновыборочный тест, который проверяет равенство дисперсии непрерывной случайной величины заданному порогу. В Python для него нет готовой функции.

сурс: https://habr.com/ru/companies/mygames/articles/677074/
Про третий говорить не будем, а про первые два поговорим.
Самый распространённый, самый известный – первый. С него и начнём - проверка выборка-теория:

К сожалению, если вы пойдёте гуглить, то в 99% источников это сведётся к «да/нет» - «гипотеза верна/неверна». На самом деле этот тест гораздо более инофрмативный: он позволяет измерить конкретное значение вероятности соответствия/несоответствия статистики теоретическому распределению. Просто если эта вероятность оказалась 92%, а мы задрали планку доверия в 95%, то пишем «гипотеза неверна», а если планка всего лишь 90%, то «гипотеза верна». Короче, вся эта хрень с доверительными интервалами высосана из пальца.

Но как этот критерий вычисляет вероятность соответствия?
Для начала вспомним, что такое хи-квадрат распределение. Это когда мы k раз проделываем какие-то измерения. Глянем Википедию:

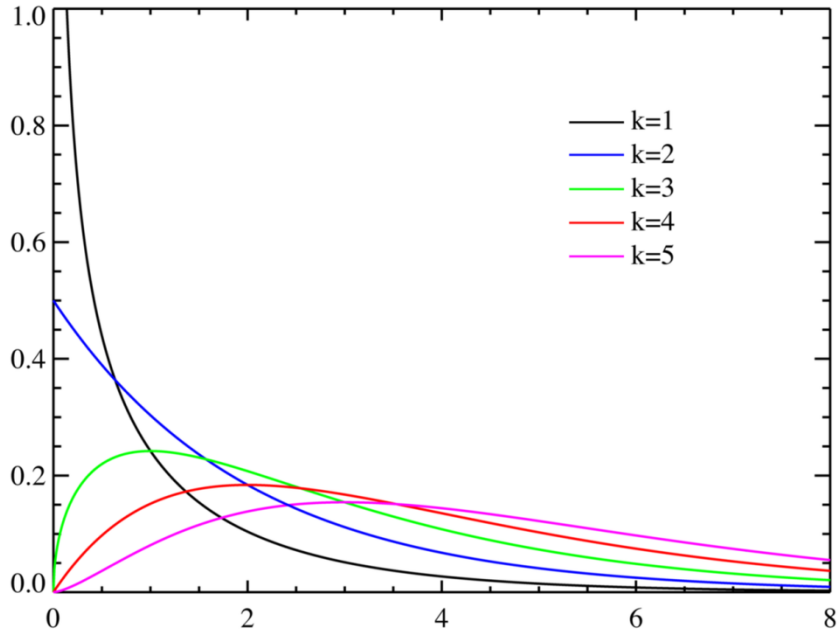
Пусть $z_1, \ldots, z_k$ — совместно независимые стандартные нормальные случайные величины, то есть: $z_i \sim N(0,1)$. Тогда случайная величина

$$x = z_1^2 + \ldots + z_k^2$$

имеет распределение хи-квадрат с k степенями свободы, то есть $x \sim f_{\chi^2(k)}(x)$, или, если записать по-другому:

$$x = \sum_{i=1}^{k} z_i^2 \sim \chi^2(k).$$

Обратите внимание, что дисперсия всех измерений 1. Давайте считать, что в нашем эксперименте так оно и есть. Тогда плотность вероятности хи-квадрата – суммарной дисперсии есть



Чем больше измерений, тем больше суммарная дисперсия (логично).

Ах, если бы Вася-экспериментатор промерил кучу хи-квадратов – тогда на гистограмме он увидит именно хи-квадрат. Но у нас на руках есть одно-единственное измерение хи-квадрата – экспериментальное:

$$\chi^2_{\text{эксп,Васи}} = \sum_{i=1}^{k} (\text{Изм}_i - \text{МатО}_i)^2$$

k - число измерений, $\text{Изм}_i - i -$ тое измерение, МатО – матожидание $i -$ того измерения. Оно ищется НЕ как экспериментальное матожидание (среднее точек выборки), а из заранее готовой теормодели, предложенной Василием для сверки с выборкой.

А далее Васе нужно по этой ЕДИНСТВЕННОЙ точке понять, норм или не норм ☺

Идея хи-квадрат теста на соответствие распределении очень проста:

Маленькая погрешность (суммарный хи-квадрат) – годится!

Большая погрешность (суммарный хи-квадрат) – не годится!

Вася накладывает подсчитанный им хи-квадрат на график теоретического распределения:

Который делит площадь под графиком на 2 части. Суммарная площадь этих кусков, 1 и 2 равна 1. А вот площадь конкретно куска 2 и есть вероятность того, что эксперимент соответствует теории.

Т.е. вероятность соответствия эксперименту теории есть

$$\int_{\chi^2_{\text{эксп}}}^{+\infty} \chi^2_k(x)dx$$

Чуть понятней, но объёмистей будет записать

$$\int_{\chi^2_{\text{эксп}}}^{+\infty} \text{плотность вероятности } \chi^2_k(\text{величина размерности } \chi^2)d\text{величина размерности } \chi^2$$

В большинстве источников, повторюсь, вместо подсчёта вероятности начинаются игры с квантилями и уровнями значимости в духе

А) нужно с потолка взять «границу выбросов», т.е. квантиль доверия, соответствующий тому, насколько крут научный журнал, куда вы хотите отправить статью
Б) посмотреть, куда попал хи-квадрат Васи
В) сделать выбор да/нет.

Обсудим ещё некоторые нюансы.
0) Хи-квадрат исповедует принцип «я его слепила из того, что было». Вася может, например, проведя 50 измерений и проведя хи-квадрат-тест для них, провести хи-квадрат-тест и для 10 первых измерений (взяв, конечно, уже другую диффеенциальную функцию распределения: $\chi^2_{10}(\chi^2)$, а не $\chi^2_{50}(\chi^2)$. Да хоть для трёх измерений – критерий Пирсона сработает всегда! (А вот верить ли ему – решать вам).

1) Если вы будете гуглить, то наткнёсь на формулу $\sum_i \frac{(O_i - E_i)^2}{E_i}$
где $O_i -$ измерение $(obvious), E_i -$ матожидание $(expected)$. Немного другие обозначения. Я такие не люблю – всё-таки сложно свыкнуться с тем, что ожидание – это не O_i.

2) Как вы видите из $\sum_i \frac{(O_i - E_i)^2}{E_i}$, тут есть матожидание в знаменателе. А вот это мне непонятно, что это и зачем. Особенно это весело, когда матожидание равно 0. И даже если не 0! Представим себе, что мы пытаемся доказать, что точка движется равномерно вдоль оси x со скоростью 1:

Время	Координата

0	0,4
1	1,4
2	2,3
3	3,5
4	4,5
5	5,4

Но стоит нам сдвинуть ось x:

Время	Координата
0	0,1
1	1,1
2	2,0
3	3,2
4	4,2
5	5,1

И тут же хи-квадрат поразительно подскачет, хотя по физике ничего не поменялось. В общем, с этим знаменателем бред полный.

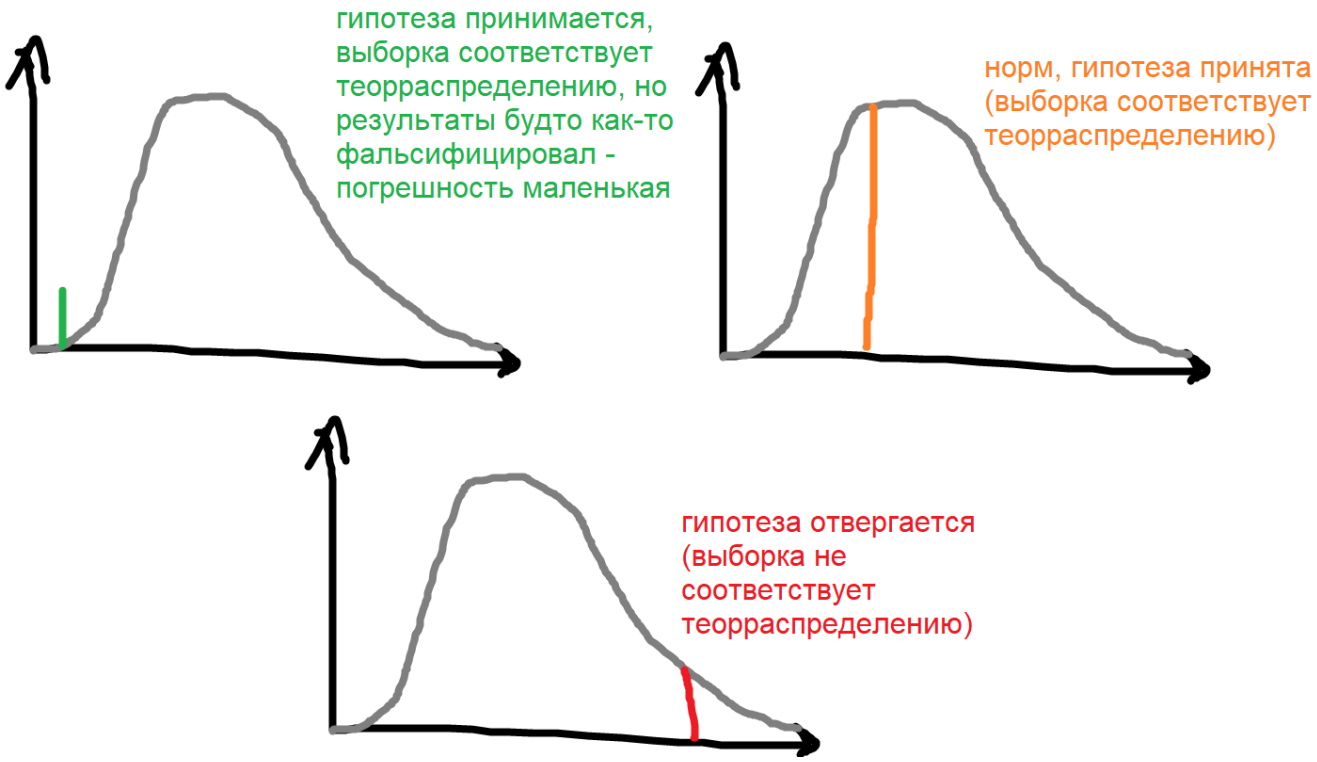
3) Кстати, вернёмся к этому примеру с равномерным движением. Для матожидания нам нужно знать теоретическое предсказание в каждой точке:

Время	Координата	Матожидание	Дисперсия от точки
0	0,4	?	?
1	1,4	?	?
2	2,3	?	?
3	3,5	?	?
4	4,5	?	?
5	5,4	?	?

Т.е. точную начальную координату и точную скорость. Их откуда брать?
А вот это метод Пирсона не интересует. Находите как хотите – методом наибольшего правдоподобия, угадывайте, из других измерений. И лишь когда вы так или иначе 

(теоретических параметров θ_i) - вы можете проверить правильность сделанного вами выбора критерием Пирсона.
В частности, если вы до этого очень серьёзно подбирали теоретическую модель-предсказание (возможно, даже методом максимального правдоподобия), а после применения Пирсона у вас гипотеза о выполнении на уровне 20%, это означает, что ~~установка говно~~ у вас кривые руки. А если теоретическая модель была сделана на коленке, то 20% от Пирсона ещё не о чём не говорят, кроме того, что ваша модель с коленки – говно.

4) В формуле есть нюанс, связанный с числом степеней свободы. В общем, нужно не k, а k-1. Не спрашивайте, почему. Карл Пирсон, кстати, тоже сам до этого не додумался. Это уже потом другой статистик, Фишер, сообразил.
5) Иногда в хи-квадрат тесте выделяют ещё вот такой случай, который отмечен зелёным:



Когда итоговый экспериментальный хи-квадрат оказывается уж слишком маленьким, меньше пика.
Гипотеза в этом случае всё равно принимается (куда уж лучше – минимальная погрешность), но «осадочек» остаётся: повод подойти к лаборантам и спросить

Теперь перейдём ко второму хи-квадрат тесту:

2. Тест на независимость (он же test of independence/association) — непараметрический, одновыборочный тест, который проверяет наличие связи между двумя категориальными переменными. В Python реализован функцией `scipy.stats.chi2_contingency`.

Как правило, он иллюстрируется на примере доказательства эффективности лекарства. Формула уже чуть другая:
Сумма становится двойной. k и К – число измерений в каждой из выборок.

$$\sum_{i=1}^{k}\sum_{I=1}^{K}(\text{Изм}_i - \text{МатО}_i)^2$$

Далее снова ищем эти точку в хи-квадрат распределении с числом степеней уже (k-1)(K-1). А в остальном то же самое.

Другие тесты

Одним распределением хи-квадрат жизнь не ограничится. Рассмотрим интереса ради ещё одно распределение – распределение Фишера.
Пусть имеются две выборки объёмом m и n соответственно случайных величин X и Y, имеющих нормальное распределение. Необходимо проверить равенство их дисперсий.
Считаем отношение экспериментальных дисперсий выборок:

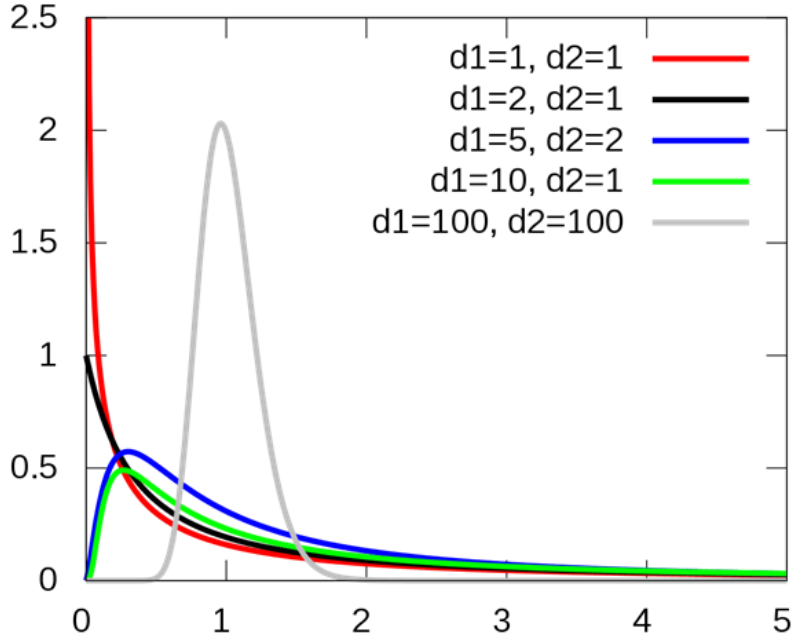
$$F_{\text{эксп}} = \frac{\sigma_X^{\text{эксп}}}{\sigma_Y^{\text{эксп}}}$$

и ищем эту точку в двухпараметрическом распределении Фишера (как вы видите, оно уже двухпараметрическом):

Пусть Y_1, Y_2 — две независимые случайные величины, имеющие распределение хи-квадрат: $Y_i \sim \chi^2(d_i)$, где $d_i \in \mathbb{N},\ i = 1, 2$. Тогда распределение случайной величины

$F = \dfrac{Y_1/d_1}{Y_2/d_2}$ называется распределением Фишера (распределением Снедекора) со степенями свободы d_1 и d_2. Пишут $F \sim \mathrm{F}(d_1, d_2)$.

Выглядит оно уже так (а по уму график в 3D надо рисовать):



Правда, в случае критерия Фишера возникнет проблема со взятием интеграла от точки до $+\infty$, как это было с хи-квадратом: $\int_{\chi^2_{\text{эксп}}}^{+\infty}$ ..., т.к. у нас уже будет двумерная область интегрирования. (Честно, не знаю, как это решается).