F-тест (критерий Фишера)

В математической статистике обычно рассматривается нулевая гипотеза H_0, которая считается верной по умолчанию, и альтернативная ей гипотеза H_1. Основная задача — ответить на вопрос: согласуются ли измерения с нулевой гипотезой, или стоит нам отвергнуть её?

Чтобы ответить на этот вопрос, вводят понятие *статистики*. Пусть у нас есть выборка из N измерений $\vec{x} = (x_1, \ldots, x_N)$ некоторой случайной величины ξ с плотностью распределения $f(\xi)$. Вообще говоря функция распределения $f(\xi)$ различная при разных гипотезах: $f(\xi|H_0)$ и $f(\xi|H_1)$. *Статистикой* называется скалярная функция $t(\vec{x})$, которая принимает в качестве аргумента наблюдаемую выборку $\vec{x}$. Есть большая свобода в выборе статистики t, и она подбирается под определённую задачу. Как видно, статистика является случайной величиной и, поэтому имеет свою функцию распределения при различных гипотезах $g(t|H_0)$ и $g(t|H_1)$.

Часто статистику строят так, что при альтернативной гипотезе H_1 значения t принимают большие значения, нежели при H_0. *Критическим значением* и *критическим регионом* называется t_c и соответственно $t > t_c$ — это регион, на котором при таких значениях статистики мы отвергаем нулевую гипотезу H_0. Регион определяется *уровнем значимости* α — вероятность ошибочно отвергнуть нулевую гипотезу, когда она верна. Уровень значимости задаётся нами, и обычно его принимают равным 5% ($\alpha = 0,05$)

$$\alpha = \int_{t_c}^{\infty} g(t|H_0)dt$$

Также важной величиной является *мощность теста* $1 - \beta$. Полезность теста определяется в его способности различать нулевую гипотезу и альтернативную гипотезу: чем более разделены H_0 и H_1, тем больше мощность теста $1 - \beta$.

$$\beta = \int_{-\infty}^{t_c} g(t|H_1)dt$$

p-значение — это вероятность получить такое же или более экстремальное значение статистики по сравнению с ранее наблюдаемым t_{obs}, при условии, что нулевая гипотеза верна. p-значение является мерой наблюдаемого уровня значимости. Оно является функцией данных и, следовательно, случайной величиной. Его не следует путать с уровнем значимости α, который представляет собой заранее определенную константу. Если мы выбрали $\alpha = 0,05$ и при наблюдении обнаружили, что $p_0 < 0,05$, то результат считается статистически значимым, позволяя отклонить нулевую гипотезу.

$$p_0 = \int_{t_{obs}}^{\infty} g(t|H_0)dt$$

1 Критерий Колмогорова-Смирнова

Рассмотрим пример широко известного и универсального теста — критерий Колмогорова-Смирнова. Это непараметрический тест гипотез, который измеряет вероятность

того, что выбранная одномерная выборка данных получена из той же генеральной совокупности, что и вторая выборка (the two-sample KS test), либо из непрерывной теоретической модели (the one-sample KS test). В данном тесте нулевой гипотезой H_0 является "две выборки имеют одинаковое распределение" или, например, "выборка порождена нормальным распределением".

Тест основан на статистике D, которая измеряет наибольшее расстояние между кумулятивными функциями распределения (CDF).

$$D = \max_x |F_1(x) - F_2(x)| \quad \forall\, x$$

Чтобы построить CDF на основе наблюдения, для одной из выборок используют эмпирическую функцию наблюдения (EDF):

$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} I(x_i \leq x)$$

$$I(x_i \leq x) = \begin{cases} 1, & x_i \leq x, \\ 0, & x_i > x. \end{cases}$$

Если нужно учесть ошибки измерений, то легче всего использовать метод Монте-Карло или bootstrap.

Во многих отношениях тест Колмогорова–Смирнова кажется очень привлекательным для использования. Его преимущества включают:

1. Тест является непараметрическим и не зависит от конкретного распределения (при условии непрерывности распределения генеральной совокупности или модели). Это означает, что он даёт корректные вероятности для любого исходного распределения данных и сравниваемой выборки. Это особенно важно, поскольку математическое распределение изучаемых объектов обычно неизвестно.

2. Одновыборочный KS-тест может использоваться как тест согласия (goodness-of-fit) после регрессии или другой статистической процедуры.

3. Тест можно применять практически к любой научной задаче — нет строгих ограничений на размер выборки. Критические значения широко доступны:

   - существуют асимптотические формулы для больших выборок (примерно $n > 30$),
   - а также табличные значения для малых выборок.

Однако кажущаяся простота KS-теста может вводить в заблуждение. Во-первых, даже когда тест корректно применим, он часто недостаточно чувствителен для выявления различий между распределениями, тогда как другие тесты, основанные на эмпирической функции распределения (EDF), показывают лучшую эффективность. Кроме того, иногда KS-тест применяют в ситуациях, для которых он не предназначен, что приводит к неправильным вероятностям при проверке гипотез.

2 Тест Андерсона–Дарлинга

KS-тест наиболее чувствителен, когда эмпирические функции распределения (EDF) отличаются глобально, особенно в центре распределения.

Однако если различия между EDF повторяются многократно, или EDF имеют (или подгоняются так, чтобы иметь) одинаковые средние значения, то функции распределения пересекаются несколько раз, и максимальное отклонение между ними уменьшается.

Тест Крамера–фон Мизеса (CvM), который измеряет сумму квадратов отклонений между EDF, лучше справляется с такими случаями. Но и KS, и CvM мало чувствительны, когда различия между кривыми наиболее заметны в начале или в конце распределения. Это происходит потому, что по определению EDF стремятся к 0 и 1 на концах распределения, поэтому отклонения там всегда небольшие.

Для решения этих проблем в 1950-х годах был разработан тест Андерсона–Дарлинга (AD) — взвешенная версия теста CvM. В многочисленных исследованиях показано, что он всегда более чувствителен, чем KS-тест. Он также непараметрический, не зависит от распределения, поэтому его можно применять практически всегда. Однако распределение статистики AD для малых выборок довольно сложное, и вычислительные алгоритмы появились лишь относительно недавно.

KS-тест нельзя применять в двух и более измерениях

Часто данные представляют собой точки на плоскости или в многомерном пространстве, а не на одной оси. В некоторой литературе можно встретить описание двумерного KS-теста. Однако ни один тест на основе EDF (включая KS, AD и аналогичные) не может быть корректно применён в двух и более измерениях.

Причина в том, что не существует уникального способа упорядочить точки, чтобы корректно вычислить расстояние между эмпирическими функциями распределения. Можно построить некоторую статистику на основе определённого порядка точек и затем вычислить максимальное расстояние между выборками (или между выборкой и кривой). Но критические значения такой статистики уже не будут независимы от распределения.

Здесь может помочь bootstrap, позволяющий численно оценить уровни значимости для конкретной статистики и конкретного многомерного набора данных.