Глава 8

Проверка статистических гипотез

Пусть известна выборка из N независимых реализаций некоторой случайной величины $x_1, ..., x_N$, для краткости обозначаемая вектором $\mathbf{x}$. Предположим, что мы хотим вынести некоторое суждение о распределении этой случайной величины. Например, убедиться, что случайная величина распределена нормально, удостовериться, что случайная величина обладает нулевым средним, или проверить, что две независимые случайные величины имеют одинаковое среднее. В простейших случаях, например, когда доступная выборка имеет очень большой размер, ответ на эти вопросы может оказаться очевидным, например, после построения гистограммы или квантиль-квантильного графика. Однако, в тех случаях, когда ответ не очевиден, мы хотели бы иметь формальный математический аппарат, позволяющий вынести аргументированное суждение о статистических свойствах.

Итак, *проверка статистических гипотез* заключается в том, что нужно сделать выбор между двумя утверждениями, называемыми обычно *нулевой гипотезой* H_0 и *альтернативной гипотезой* H_1. Например:

- Нулевая гипотеза H_0: $p(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right)$.
- Альтернативная гипотеза H_1: $p(x_i) = \frac{1}{4} \exp\left(-\frac{|x_i|}{2}\right)$.

Утверждения вида $p(x_i) = p_0(x_i)$ (следовательно $p(\mathbf{x}) = \prod_{i=1}^{N} p_0(x_i)$) часто называют *простой нулевой гипотезой*, а утверждения $p(x_i) = p_1(x_i)$ — *простой альтернативной*, смысл здесь в том, что параметры распределения считаются известными и не подлежат определению. В противном случае, гипотезу называют *сложной*. Например, если $p(x_i) = p'_0(x_i|\theta)$ известно с точностью до неизвестных параметров θ, которые подлежат оцениванию в рамках проверки статистической гипотезы.

С формальной точки зрения у нас в руках есть только выборка независимых реализаций случайной величины, записываемая как вектор $\mathbf{x} \in \mathrm{R}^N$.

И фактически, мы ищем способ разбить пространство R^N на две части. При попадании конкретной выборки, обозначаемой вектором $\mathbf{x}$ в одну из которых, называемую *критической областью* C, нулевая гипотеза отвергается, и, соответственно, при попадании в другую, нулевая гипотеза принимается. Кстати, в этом задача проверки статистических гипотез имеет некоторое сходство и параллели с задачей бинарной классификации из области машинного обучения.

С практической точки зрения проще всего оказывается ввести некоторую новую действительную функцию $f(\mathbf{x})$, которая затем сравнивается с некоторым пороговым значением K, и, например, если $f(\mathbf{x}) > K$, то нулевая гипотеза отклоняется, и наоборот. Функция $f(\mathbf{x})$ и порог K определяются некоторым оптимальным образом для каждой конкретной задачи, собственно, данная глава посвящена примерам выбора статистических критериев.

Основная сложность состоит в том, что чаще всего нулевая гипотеза допускает, что реализация $\mathbf{x}$ может принимать любые значения в R^N с ненулевой вероятностью, это значит, что даже при выполнении нулевой гипотезы вектор $\mathbf{x}$ может случайно оказаться в критической области, и нулевая гипотеза будет ошибочно отвергнута. Неравноправие в терминологии (нулевая и альтернативная гипотезы) во многом связано с понятием ошибок первого и второго рода:

- *Ошибкой первого рода* называется ситуация, когда гипотеза H_0 на самом деле верна, но была отклонена в рамках проверки.
- *Ошибкой второго рода* называется ситуация, когда гипотеза H_1 на самом деле верна, но была принята гипотеза H_0.

Понятно, что стоимость последствий ошибок первого и второго рода различна, и определяется исходя из внешних соображений. Например, допустим что анализируются данные некоторого медицинского исследования, и проверяются две гипотезы. Нулевая гипотеза H_0 состоит в том, что у пациента имеется некоторое очень опасное заболевание, а альтернативная H_1 — в том, что пациент здоров. Последствия ошибки первого рода будут состоять в том, что лечение пациента не будет начато своевременно, что приведет к непоправимым последствиям для его здоровья, или даже летальному исходу. В то время как последствия ошибки второго рода будут состоять в том, что пациенту придется потратить время, пройти ряд дополнительных исследований, возможно, поволноваться, но конечном счете выяснится, что ему ничего не угрожает. В данном случае, очевидно, что последствия ошибки первого рода для данного человека будут иметь более высокую стоимость.

Итак, в рамках проверки статистических гипотез нам хотелось бы так задать критическую область C, чтобы вероятность ошибки первого рода не превосходила некоторого приемлемого для нас порога α, называемого *уровнем значимости*:

$$P\left\{\text{Ошибка первого рода}\right\} = \int_C p_0(\mathbf{x})d\mathbf{x} \leq \alpha. \tag{8.1}$$

Заметим, что существует вырожденный (и не интересный) случай, когда всегда принимается основная гипотеза, поэтому на практике мы фиксируем ошибку первого рода на ее верхнем допустимом уровне α.

Вероятность ошибки второго рода должна быть сделана по возможности минимальной:

$$P\{\text{Ошибка второго рода}\} = \int_{\bar{C}} p_1(\mathbf{x})d\mathbf{x} = 1 - \int_C p_1(\mathbf{x})d\mathbf{x}. \tag{8.2}$$

Интеграл $\int_C p_1(\mathbf{x})d\mathbf{x}$ часто называют *мощностью критерия*.

На практике, как упоминалось, большинство критериев строится с помощью статистики $f(\mathbf{x})$ и порога K, в этом случае удается определить параметрическое семейство критических областей $C(K)$ и выбрать среди них критическую область для требуемого уровня значимости α. Иногда дополнительно удается доказать, что предложенный критерий наиболее мощный, т.е. имеет минимальную ошибку второго рода при заданном уровне значимости α.

8.1 Критерий Неймана-Пирсона

Рассмотрим отношение правдоподобий

$$\Lambda_{H_1,H_0} \equiv \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} \tag{8.3}$$

Гипотеза H_0 отвергается в том случае, когда $\Lambda_{H_1,H_0} \geq K$, некоторого порогового значения, которое по сути является функцией уровня значимости $K = K(\alpha)$. Лемма Неймана-Пирсона состоит в том, что данный критерий является наиболее мощным (минимальная вероятность ошибки второго рода), среди всех критериев с уровнем значимости (вероятностью ошибки первого рода) α.

Рассмотрим следующий пример из книги Кельберт и Сухов 2017. Пусть, для простоты рассмотрения, у нас есть одно измерение x_1, и мы хотим найти критерий отношения правдоподобий с уровнем значимости $\alpha = 0.05$. Тогда построим отношение правдоподобий:

$$\Lambda_{H_1,H_0} = \frac{\sqrt{2\pi}}{4} \frac{\exp\left(-\frac{|x_1|}{2}\right)}{\exp\left(-\frac{x_1^2}{2}\right)}. \tag{8.4}$$

Понятно, что $\Lambda_{H_1,H_0} \geq K(\alpha)$ эквивалентно утверждению $x_1^2 - |x_1| \geq K'(\alpha)$, таким образом критическая область $C(\alpha)$ в нашем случае задается двумя полупрямыми:

$$|x_1| \geq \frac{1}{2} + t(\alpha), \tag{8.5}$$

$$|x_1| \leq \frac{1}{2} - t(\alpha). \tag{8.6}$$

Рассчитаем вероятность ошибки первого рода, чтобы найти конкретное значение порога $t(\alpha)$.

Если $t \geq \frac{1}{2}$, тогда

$$\alpha = \int_C p_0(x_1)dx_1 = \\ = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{-\frac{1}{2}-t} \exp\left(-\frac{x_1^2}{2}\right) dx_1 + \int_{\frac{1}{2}+t}^{\infty} \exp\left(-\frac{x_1^2}{2}\right) dx_1 \right) = \\ = 1 - \mathrm{erf}\left(\frac{1}{2} + t\right), \quad (8.7)$$

откуда следует, что $t = \mathrm{erf}^{-1}(1-\alpha) - \frac{1}{2}$, и для $\alpha = 0.05$ получается критическая область

$$|x_1| \geq 1.38\ldots . \quad (8.8)$$

Если $t \leq \frac{1}{2}$, тогда

$$\alpha = 1 - \mathrm{erf}\left(\frac{1}{2} + t\right) + \mathrm{erf}\left(\frac{1}{2} - t\right), \quad (8.9)$$

и это уравнение не имеет решений для выбранного $\alpha = 0.05$.

В данном случае $|x_1|$ играет роль статистики критерия, и нам остается сравнить эту величину с пороговым значением, специально найденным так, чтобы гарантировать выбранный уровень значимости $\alpha = 0.05$.

8.2 Критерий Стьюдента

Рассмотрим другую задачу. Пусть известно N независимых реализаций некоторой случайной величины $x_1, ..., x_N$. Гипотеза H_0 состоит в том, что случайная величина распределена согласно нормальному закону с средним μ_0, гипотеза H_1 состоит в том, что случайная величина распределена согласно нормальному закону с некоторым другим неизвестным средним $\mu \neq \mu_0$. Подразумевается, что величина μ_0 может предсказывается нам какой-то теорией, и мы хотели бы проверить или опровергнуть теорию на основе измерений.

Понятно, что почти невероятно, что для конечного числа N выборочное среднее

$$m \equiv \frac{1}{N} \sum_{i=1}^{N} x_i, \quad (8.10)$$

в точности совпадет с величиной μ_0, поэтому нужен критерий допускающий отклонение выборочного среднего m от ожидаемого μ_0 в разумных приделах.

Рассмотрим величину

$$T \equiv \frac{m - \mu_0}{\sqrt{\frac{\sum_i^N (x_i - m)^2}{N(N-1)}}}, \quad (8.11)$$

которая имеет смысл отношения отклонения выборочного среднего от μ_0 к ошибке выборочной оценки среднего. Иначе говоря, это относительное отклонение, заданное в единицах ошибки выборочной оценки среднего. Интуитивно понятно, что чем больше это относительно отклонение, тем менее вероятно равенство $m = \mu_0$.

Британский ученый Вильям Госсет (известный под псевдонимом Стьюдент) показал, что при выполнении H_0 такая величина T распределена в соответствии с распределением Стьюдента $p_{N-1}(T)$ с $N-1$ степенью свободы. Напомним, что плотность вероятности распределения Стьюдента задаётся следующим образом:

$$p_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{8.12}$$

где целочисленный параметр ν называют числом степеней свободы распределения.

Значит можно вычислить вероятность ошибки первого рода для критерия $|T| > a$:

$$\alpha = \int_{-\infty}^{-a} p_{N-1}(T)dT + \int_{a}^{\infty} p_{N-1}(T)dT = 2\gamma_{N-1}(a). \tag{8.13}$$

К несчастью, интеграл не выражается в элементарных функциях, поэтому обозначим кумулятивную функцию распределения Стьюдента как $\gamma_{N-1}(a)$, тогда решение уравнения записывается через обратную функцию, называемую квантильной функцией, как $a = t_{N-1}(\frac{\alpha}{2})$. Итак, если $|T| \geq t_{N-1}\left(\frac{\alpha}{2}\right)$, то гипотеза H_0 отклоняется на уровне значимости α. Конкретное значение функции $t_{N-1}\left(\frac{\alpha}{2}\right)$ может быть найдено численно с помощью готовых программных пакетов.

Полученный результат часто формулируют в терминах доверительных интервалов (т.е. интервал, который накрывает настоящее значение μ с вероятностью $1-\alpha$):

$$P\left\{m - \sqrt{\frac{\sum_i^N (x_i-m)^2}{N(N-1)}}t_{N-1}\left(\frac{\alpha}{2}\right) < \mu < m + \sqrt{\frac{\sum_i^N (x_i-m)^2}{N(N-1)}}t_{N-1}\left(\frac{\alpha}{2}\right)\right\} = 1-\alpha. \tag{8.14}$$

Для проверки статистических гипотез часто вместо порога $t_{N-1}(\frac{\alpha}{2})$ используется так называемое p-значение[1], определяемое как значение кумулятивной функции распределения от величины статистики, в случае критерия Стьюдента $p = \gamma_{N-1}(T)$. Благодаря монотонности кумулятивной функции любого распределения, критерий Стьюдента теперь записывается в виде $p(T) \geq \alpha$. Такая система единиц позволяет использовать единый формализм для многих статистических критериев, ведь величина p принимает значения на интервале $[0,1]$, в отличии от порога, который имеет индивидуальные характерные значения для каждого критерия.

[1] p-value

8.3 Критерий Пирсона

Пусть известно N независимых реализаций некоторой случайной величины $x_1, ..., x_N$. Нулевая гипотеза H_0 состоит в том, что случайная величина распределена с известным распределением $p_0(x)$.

Разобьем область значений x на K не пересекающихся интервалов, и для каждого интервала подсчитаем количество попаданий N_l в этот интервал. При нулевой гипотезе среднее число попаданий будет

$$e_l = Np_l = N\int_{D_l} p_0(x)dx, \tag{8.15}$$

и может быть рассчитано тем или иным образом для выбранного разбиения D_l и $p_0(x)$. При альтернативной гипотезе число попаданий будет произвольной величиной.

Рассмотрим величину

$$P_N \equiv \sum_{l=1}^{K} \frac{(N_l - e_l)^2}{e_l}. \tag{8.16}$$

Она имеет смысл взвешенной суммы квадратов отклонений числа попаданий в интервал от теоретических средних значений.

Британский ученый Карл Пирсон доказал, что

$$\lim_{N\to\infty} P\{P_N > \alpha\} = \int_{\alpha}^{\infty} p_{N-1}(x)dx, \tag{8.17}$$

где плотность вероятности χ^2-распределения Пирсона с k степенями свободы задаётся выражением:

$$p_k(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{k/2-1} \exp\left(-\frac{x}{2}\right), \tag{8.18}$$

где целочисленный параметр k называется числом степеней свободы.

Соответственно, гипотеза H_0 отклоняется с уровнем значимости α, если $P_N > \int_{\alpha}^{\infty} p_{\chi^2_{N-1}}(x)dx$.

8.4 Критерий Колмогорова-Смирнова

Рассмотрим предыдущую задачу, но предложим другой способ её решения. Построим эмпирическую кумулятивную функцию распределения:

$$F_N(\mathbf{x}) \equiv \frac{1}{N}\sum_{i=1}^{N} I_{(-\infty,x]}(x_i), \tag{8.19}$$

где $I_{(-\infty,x]}$ обозначает индикаторную функцию, равную 1 для $x_i < x$ и нулю в противном случае.

Теоретическую кумулятивную функцию распределения при нулевой гипотезе мы знаем:

$$F(x) = \int_{-\infty}^{x} p_0(x)dx. \tag{8.20}$$

Рассматривается величина

$$D_N \equiv \sup_x |F(x) - F_N(x)|, \tag{8.21}$$

которая имеет смысл максимального отклонения между эмпирической и теоретической кумулятивными функциями. Заметим, что при $x = \pm\infty$ эти две функции всегда совпадают по построению, значит максимальное отклонение достигается где-то между этими точками.

Советский математик Андрей Николаевич Колмогоров изучил свойства распределения величины K, определяемой как:

$$K \equiv \sup_{t\in[0,1]} |B(t)|, \tag{8.22}$$

где $B(t)$ — Броуновский мост (про Броуновский мост см., например, Степанов 2012), случайный нестационарный процесс, для которого, в частности, $p(0, t = 0) = p(0, t = 1) = 1$. Полученное распределение называют распределением Колмогорова, оно показывает насколько сильно по амплитуде отклоняется Броуновский мост. Плотность вероятности и кумулятивная функция такого распределения задается в виде бесконечного ряда, и вычисляются с помощью численных методов.

Оказывается, что величина $\sqrt{N}D_N$ стремится к $\sup_x |B(F(x))|$ по распределению при $N \to \infty$. Таким образом, если $\sqrt{N}D_N > K(\alpha)$, то нулевая гипотеза отвергается, где $K(\alpha)$ находится из квантильной функции распределения Колмогорова для требуемого уровня значимости α. Асимптотически, мощность критерия стремится к единице, это значит, что он не совершает ошибок второго рода.

При попытке обобщить критерий Колмогорова-Смирнова на большие размерности случайных величин возникает интересное препятствие — ответ критерия становится не инвариантен относительно аффинных преобразований, т.е. простая линейная замена переменных влияет на результат. Поэтому, например, для проверки принадлежности выборки векторов к многомерному нормальному распределению существуют отдельные критерии, построенные исходя из принципа инвариантности относительно аффинных преобразований. Обзор различных критериев приведен, например, в работе Henze 2002.