Критерии согласия и их эффективность

Ragozin Ilya
Высшая школа экономики, Санкт-Петербург, Россия

26 января 2023 г.

Непараметрическая статистика

Непараметрическая статистика – раздел статистики, который не основан исключительно на параметризованных семействах вероятностных распределений.
Рассмотрим некую выборку:

$$X_1, X_2, ..., X_n \sim F(\theta),\ \theta \in \mathbb{R}^d\ : \theta = (\theta_1, ..., \theta_d).$$

Рассмотрим следующие возможные гипотезы:
1. $H_0\ :\ \theta_1 = a;$
2. $H_0\ :\ \theta_1 = \theta_2;$
3. $H_0\ :\ \theta = \theta_0;$
4. $H_0\ :\ F = F_0,\ \theta = \theta_0;$
5. $H_0\ :\ F = F_0;$
6. $H_0\ :\ F \in \mathfrak{F};$

Примеры. Критерий Пирсона.

Один из самых важный примеров критерия согласия это критерий согласия Пирсона, основанный на следующей теореме:
Общие положения: $X_1, .., X_n \sim F(X)$; $H_0 : F(x) = F_\theta(x)$; где $\theta = (\theta_1, ..., \theta_r)$.
Теорема:

$$\chi_n^2 = \sum_{i=1}^{k} \frac{(v_i - np_i)^2}{np_i} \to \chi_{k-r-1}^2;$$

где k-число интервалов, в которые сгруппированы случайные величины (Δ_k) и v_k –соответствующая частота попадания выборки в заданный интервал;
$p_i = P(X \in \Delta_i)$, где $P(\cdot)$ определяется распределением из H_0.

Другие идеи для построения критериев согласия

Первой и довольно логичной идеей было использовать эмпирическую функцию распределения для выборки $X_1, ..., X_n \sim F$:

$$F_n(t) = \sum_{i=1}^{n} Ind\{X_i < t\};$$

Используя свойства э.ф.р., а именно $F_n(t)$– состоятельная, несмещенная, асимптотическая нормальная оценка функции распределения F. Тогда вполне логично рассмотрение статистик, основанных на разности $F_n(t) - F(t)$. Однако, к сожалению, долгое время это не представлялось возможным в теоретическом смысле, так как не было механизма, позволяющего исследовать такого вида статистики. Привлекательность идеи заключается и в том, что по теореме Гливенко-Кантелли такая разность сходится равномерно к нулю, почти наверное, т.е.

$$\sup_{t \in R} |F_n(t) - F(t)| \to 0,\ n \to \infty.$$

Другие идеи для построения критериев согласия

Характеризация–такое свойство для определенного вида распределений, которое выполняется только для него, посредством одинаковой распределенности некоторых статистик, н.п. :

$$g_1(X_1, ..., X_d) \stackrel{d}{=} g_2(X_1, .., X_s) \Leftrightarrow X \sim F(x) \in \mathfrak{F};$$

Запись $\stackrel{d}{=}$ означает, что выполнено следующее равенство:
$\forall t: \ P(g_1(\cdot) < t) = P(g_2(\cdot) < t)$ – **вероятностное или характеризационное равенство**.
В 50-х годах 20-го века, появилась идея рассматривать характеризации для построения критериев согласия, но опять же все приостановилось из-за отсутствия теории и механизмов, позволяющих реализовать такой подход.
Мало того, позднее оказалось, что можно рассматривать не только конкретный класс распределений, но и более широкий класс, объединяющий в себе разные классы распределений, такое свойство принято называть "специальным свойством".

Идея характеризации и специального свойства

Используя определение характеризации, можно рассмотреть следующие эмпирические статистики, соответствующие левой и правой частям вероятностного равенства:

$$L_n(t) = (c_n^d)^{-1} \sum_{1 \le i_1 < \ldots < i_d \le n} Ind\{g_1(X_{i_1}, \ldots, X_{i_d}) < t\},\ t \in \mathbb{R};$$

$$R_n(t) = (c_n^s)^{-1} \sum_{1 \le i_1 < \ldots < i_s \le n} Ind\{g_2(X_{i_1}, \ldots, X_{i_s}) < t\},\ t \in \mathbb{R};$$

Такие статистики называются U-эмпирическими статистиками, от слова unbiased, так как очевидно, что они все являются несмещенными оценками соответственно левой и правой частей характеризационного равенства. Тогда по теореме Гливенко-Кантелли, разность таких статистик равномерно сходится к 0 при нулевой гипотезе, т.е.:

$$\sup_{t \in R} |L_n(t) - R_n(t)| \to 0,\ n \to +\infty$$

.

Как сравнивать критерии?

Немаловажным вопросом было по какому признаку оценивать критерии и сравнивать их, как в практическом смысле так и в теоретическом, для этого рассмотрим три подхода основанных на эффективности.

Определение: Пусть дана выборка $X_1, ..., X_n$ с распределением P_θ и проверяется гипотеза согласия $H_0 : \theta \in \Theta_0$ против альтернативы $H1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$. Рассмотрим последовательности статистик T_n, V_n для проверки соответствия выборки нулевой гипотезе, определим **относительную эффективность** статистики T_n по отношению к V_n на уровне значимости α и мощности β:

$$eff_{T,V}(\alpha, \beta, \theta) = \frac{N_v(\alpha, \beta, \theta)}{N_T(\alpha, \beta, \theta)};$$

где $N_T(\cdot)$ и $N_v(\cdot)$– минимальные объемы исходной выборки, для которых последовательности статистик достигают мощности $\beta \in (0, 1)$ при уровне значимости $\alpha > 0$ и альтернативным значением параметра θ.

Виды эффективностей и подходы к их вычислению

Рассмотрим самые популярные подходу к вычислению эффективностей, главной проблемой является практическая невозможность вычисления относительных эффективностей, поэтому в данной теории принято рассматривать предельные соотношения эффективностей (асимптотическое):

1. $\lim\limits_{\alpha \searrow 0} Eff_{T,V}(\alpha, \beta, \theta)$–подход Бахадура;
2. $\lim\limits_{\beta \nearrow 1} Eff_{T,V}(\alpha, \beta, \theta)$– подход Ходжеса-Лемана,
3. $\lim\limits_{\theta \to \partial\Theta_0} Eff_{T,V}(\alpha, \beta, \theta)$– подход Питмена.

Помимо этого существуют принципиально другие подходы к вычислению АОЭ, предложенные Черновым, Калленбергом, а также Хеффдингом.

Виды статистик

Вспомним ранее введенные U-эмпирические статистики, основанные на характеризационном равенстве, на основе их разности, будем рассматривать следующие виды статистик статистики:
интегрального типа:

$$I_n = \int\limits_R (L_n(t) - R_n(t))\, dF_n(t),$$

колмогоровского типа:

$$D_n = \sup_{t \in R} |L_n(t) - R_n(t)|;$$

Также в некоторых случаях бывает "взвешенные"интегральные статистики с некоторой весовой функцией w(t):

$$\int\limits_R (L_n(t) - R_n(t))\, w(t) dt$$

Бахадуровская эффективность

В качестве механизма вычисления АОЭ критериев был выбран подход Бахадура по следующим соображениям:
– Подход Питмена не применим к статистики D_n, так как она не асимптотически нормальна, а в случае интегральных статистик эффективности по Бахадуру и Питмену совпадают;
– АОЭ по Бахадуру различает статистики, равноэффективные по Питмену;
– АОЭ по Ходжесу-Леману оказывается менее удобной и содержательной по сравнению критериев, чем базадуровская АОЭ;
– развитие теории больших уклонений и асимптотика больших уклонений U–статистик и функционалов от них.

Бахадуровская теория

Введем фундаментальные понятия бахадуровской теории:
Пусть s=(X1,X2,...)–последовательность наблюдений с общим распределением P_θ. Предполагая, что $\theta \in \Theta \subset \mathbb{R}$, и проверяется нулевая гипотеза $H_0 : \theta \in \Theta_0 \subset \Theta$ против альтернативы $H1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$.
Для проверки нулевой гипотезы используем последовательность статистик $T_n(s) = T_n(X_1, ..., X_n)$, считая критическими их большие значения. Введем обозначения:

$$F_n(t;\theta) = P_\theta(s : T_n(s) < t),\ L_n(s) = 1 - inf\{F_n(T_n(s);\theta) : \theta \in \Theta_0\};$$

где величина $L_n(s)$–достигаемый уровень или P–значение. Тогда при альтернативе выполняется сходимость по P_θ–вероятности:

$$\lim_{n\to\infty} n^{-1} \ln L_n(s) = -\frac{1}{2} c_T(\theta),\ \theta \in \Theta_1,$$

Величина $c_T(\cdot)$–точный бахадуровский наклон статистик T_n.

Бахадуровская эффективность

Теорема 1. *Пусть последовательность статистик* $\{T_n\}$ *удовлетворяет следующим условиям:*

1. $T_n \to b(\theta)$ *по* $\mathbf{P}_\theta$*-вероятности,* $\theta \in \Theta_1$*, где* $-\infty < b(\theta) < \infty$*, и*
2. $\lim\limits_{n\to\infty} n^{-1} \ln \mathbf{P}_\theta(T_n \geq a) = -k(a)$ *для любого* $\theta \in \Theta_0$ *и некоторого открытого интервала* I*, где функция* k *непрерывна на множестве* $I = \{b(\theta), \theta \in \Theta_1\}$.

Тогда при всех $\theta \in \Theta_1$ *точный наклон* $c_T(\theta)$ *существует и вычисляется по формуле:*

$$c_T(\theta) = 2k(b(\theta)).$$

Бахадуровская эффективность

Теперь определим информацию Кульбака-Лейблера $K(\theta)$ между альтернативой и нулевой гипотезой $H_0,\ \theta \in \Theta_1$:

$$K(\theta, \Theta_0) = \inf\{K(P_\theta, P_{\theta_0}) : \theta_0 \in \Theta_0\}$$

; где

$$K(Q,P) = \int\limits_R \ln \frac{Q}{P} dQ, Q \ll P; \quad or \quad \infty \quad \text{в противном случае.}$$

Таким образом по неравенству Бахадура-Рагавачари:
$C_T(\theta) \leq 2K(\theta, \Theta_0)$.

$$eff_T = \lim_{\theta \to \partial\Theta_0} \frac{c_T(\theta)}{2K(\theta, \Theta_0)}.$$

Примеры новых критериев согласия. Логистичность

Характеризации:

$$X \stackrel{d}{=} \min(X, Y) + Z \tag{1}$$

$$\min(X, Y) + Z_1 \stackrel{d}{=} \max(X, Y) - Z_2 \tag{2}$$

Эти соотношения выполняются тогда и только тогда, когда: $f(x) = l(x - \theta), \theta \in \mathbb{R}$ следующего вида:

$$l(x) = \frac{e^x}{(1 + e^x)^2}. \tag{3}$$

Примеры новых критериев согласия. Логистичность

$$F_n(t) = n^{-1} \sum_{i=1}^{n} I\{X_i < t\},\ t \in \mathbb{R}.$$

построим две U–эмпирические функции распределения:

$$U_{1,n}(t) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} I\left\{\max(X_i, X_j) < t\right\},\ t \in \mathbb{R},$$

и

$$U_{2,n}(t) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} I\left\{\min(X_i, X_j) < t\right\},\ t \in \mathbb{R}.$$

Примеры новых критериев согласия. Логистичность

$$U_n^+(t) = \int_0^\infty U_{1,n}(t+s)e^{-s}ds = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \int_0^\infty I\{\max(X_i, X_j) < t+s\} e^{-s}ds$$

$$= \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \left(e^{-\max(0,\max(X_i,X_j)-t)}\right),$$

$$U_n^-(t) = \int_0^\infty U_{2,n}(t-s)e^{-s}ds = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \int_0^\infty I\{\min(X_i, X_j) < t-s\} e^{-s}ds$$

$$= \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \left(1 - e^{(\min(X_i,X_j)-t)}\right) I\{\min(X_i, X_j) < t\}$$

$$= \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \left(1 - e^{\min(0,\min(X_i,X_j)-t)}\right).$$

Примеры новых критериев согласия. Логистичность

$$IU_n = \int_{-\infty}^{\infty} \left(U_n^+(t) - U_n^-(t)\right) dF_n(t) \tag{4}$$

$$QU_n = \sup_t |U_n^+(t) - U_n^-(t)|, \tag{5}$$