Материалы региональной конференции (с участием иностранных ученых) "Вероятностные идеи в науке и философии". – Новосибирск: Ин-т философии и права СО РАН / НГУ. – С. 115-119.

Вопросы проверки адекватности вероятностных моделей по неполным выборкам[1]

Лемешко Б.Ю., Чимитова Е.В.
НГТУ, г. Новосибирск. E-mail: chim@mail.ru

Практически в любой сфере научной деятельности, связанной с регистрацией наблюдений, возникает необходимость в статистической обработке полученных данных. На основе получаемых результатов делают выводы о свойствах исследуемых объектов. При исследовании величин типа "времени жизни", например в задачах теории надежности, в медицинских или биологических исследованиях, нередко возникает задача обработки цензурированных выборок.

Определение 1. Выборка называется цензурированной (слева или справа), если область определения случайной величины разбита на два интервала, в одном из которых известны индивидуальные наблюдения, а во втором – известно лишь количество наблюдений, попавших в интервал.

Интерес к таким задачам не снижается, так как появление цензурированных выборок оказывается естественным и обычно порождается спецификой проведения экспериментов и условиями регистрации наблюдений. Достоверность результатов статистического анализа в первую очередь зависит от степени адекватности выбранной модели анализируемым данным. Поэтому обязательным этапом является проверка гипотезы о согласии имеющихся статистических данных с выбранным теоретическим распределением.

Определение 2. Гипотеза вида $H_0 : F(x) = F(x,\theta)$, где $F(x,\theta)$ – функция распределения вероятностей, с которой проверяется согласие наблюдаемой выборки $X_1, X_2, \ldots, X_n$ независимых одинаково распределенных величин, называется простой, если θ – известное значение параметра (скалярного или векторного).

Определение 3. Гипотеза вида $H_0 : \; F(x) \in \{F(x,\theta),\, \theta \in \Omega\}$ называется сложной, если в качестве неизвестного параметра θ используется его оценка $\hat{\theta}$, вычисленная по той же выборке, по которой проверяется гипотеза о согласии. Если оценка $\hat{\theta}$ вычислена по другой выборке, то гипотеза простая.

Проверка гипотезы о согласии осуществляется по следующей схеме. Для выбранного критерия вычисляется значение S^* статистики критерия S как некоторой функции от выборки и закона распределения, с которым проверятся согласие. Для используемых на практике критериев, как правило, известны предельные распределения $G(S \mid H_0)$ соответствующих статистик при условии истинности проверяемой гипотезы H_0. Гипотеза о согласии не отвергается, если $P\{S > S^*\} = \int_{S^*}^{\infty} g(S|H_0)dS > \alpha$, где α - заданный уровень значимости, $g(S|H_0)$ – плотность предельного распределения.

Для проверки согласия эмпирического распределения $F_n(x)$ с теоретическим $F(x)$ по цензурированным данным можно использовать критерии типа Реньи [1] или критерии типа Колмогорова [2].

Двусторонняя статистика критерия Реньи в случае цензурирования слева задается выражением:

$$S_R^c = \sqrt{\frac{na}{1-a}} \cdot \sup_{F(x)\ge a} \frac{|F_n(x) - F(x)|}{F(x)},$$

а в случае цензурирования справа –

$$S_R^c = \sqrt{\frac{na}{1-a}} \cdot \sup_{F(x)\le 1-a} \frac{|F_n(x) - F(x)|}{1-F(x)}.$$

где $a \in (0,1)$ – степень цензурирования. Для этой статистики при проверке простых гипотез справедливо предельное соотношение:

$$\lim_{n\to\infty} P\left\{S_R^c < S\right\} = \frac{4}{\pi}\sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left\{-\frac{(2k+1)^2\pi^2}{8S^2}\right\} = L(S). \qquad (1)$$

[1] Работа выполнена при финансовой поддержке Минобразования РФ (проект № Т 02-3.3-3356)

Статистика Колмогорова определяется выражением

$$S_K^c = \sup_M \left| F_n(x) - F(x) \right|,$$

где $M = \{x : F(x) \ge a\}$ – в случае цензурирования слева и $M = \{x : F(x) \le 1-a\}$ – при цензурировании справа, $a \in (0,1)$ – степень цензурирования. Предельное распределение статистики Колмогорова S_K^c по цензурированным данным задается соотношением

$$P\{S_K^c < S\} = \sum_{i=-\infty}^{+\infty} (-1)^i \exp(-2i^2 S^2) \cdot P\left\{ \left| X - 2iS\sqrt{\frac{a}{1-a}} \right| < \frac{S}{\sqrt{a-a^2}} \right\} = K_a^c(S). \qquad (2)$$

Предельное распределение статистики Реньи в отличие от распределения статистики Колмогорова в случае проверки простых гипотез не зависит от степени цензурирования, и по идее, таким критерием удобнее пользоваться на практике. Однако вопрос о том, насколько хорошо распределения статистик Реньи согласуются с соответствующими предельными законами при ограниченных объемах выборок, до сих пор не исследовался. Неизвестно и то, насколько быстро сходятся к своим предельным законам распределения статистики Колмогорова в случае цензурированных выборок.

Отметим еще раз, что применение перечисленных выше критериев предполагает проверку простых гипотез H_0. Однако в практике статистического анализа очень часто приходится сталкиваться с необходимостью проверки гипотез о согласии после оценивания по этой же выборке параметров предполагаемого закона распределения. К сожалению, непараметрические критерии типа Реньи и типа Колмогорова в случае проверки сложных гипотез теряют свойство “свободы от распределения”: распределения статистик становятся зависящими от вида проверяемой гипотезы (от закона, с которым проверяется согласие; от метода оценивания параметров; от того, какие именно параметры оцениваются). Распространенная ошибка, связанная с пренебрежением существующей проблемы, когда при проверке сложной гипотезы пользуются предельными распределениями, полученными для случая простой гипотезы, чаще всего приводит к необоснованному принятию нулевой гипотезы.

Целью данной работы явилось, во-первых, исследование области корректного применения критериев типа Реньи и типа Колмогорова при проверке простых гипотез и, во-вторых, построение вероятностных моделей, аппроксимирующих предельные распределения данных статистик в случае проверки сложных гипотез.

Исследования проводились с использованием методики компьютерного моделирования и анализа статистических закономерностей. Данная методика позволяет быстро и не менее точно, чем с использованием строгого математического аппарата, находить статистические закономерности.

Результаты статистического моделирования показали, что распределения статистик типа Реньи существенно зависят от степени и структуры цензурирования. Например, на рис. 1 приведены эмпирические распределения статистики Реньи для случая проверки простой гипотезы о согласии с экспоненциальным законом при $a = 0,9$.

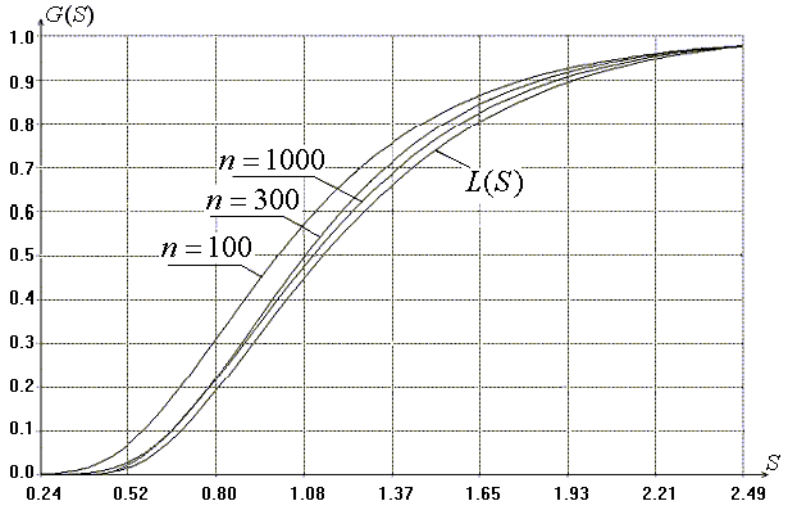

Рис. 1. Распределения статистики S_R^c, $a = 0,9$, цензурирование слева

Из рисунка видно, что относительно близкими к предельному оказываются распределения статистики при полных значениях объема выборок $n > 1000$. При меньших n распределения статистики существенно отличаются от предельного.

Рис. 2 иллюстрирует зависимость распределений статистики от степени цензурирования. На рисунке представлены распределения статистики при $n = 100$ и различной величине наблюдаемой области. Видно, что наилучшее согласие с предельным распределением достигается при 50% наблюдаемой области (при $a = 0,5$), а при малой или, наоборот, высокой степени цензурирования распределения статистики существенно отличаются от предельного.

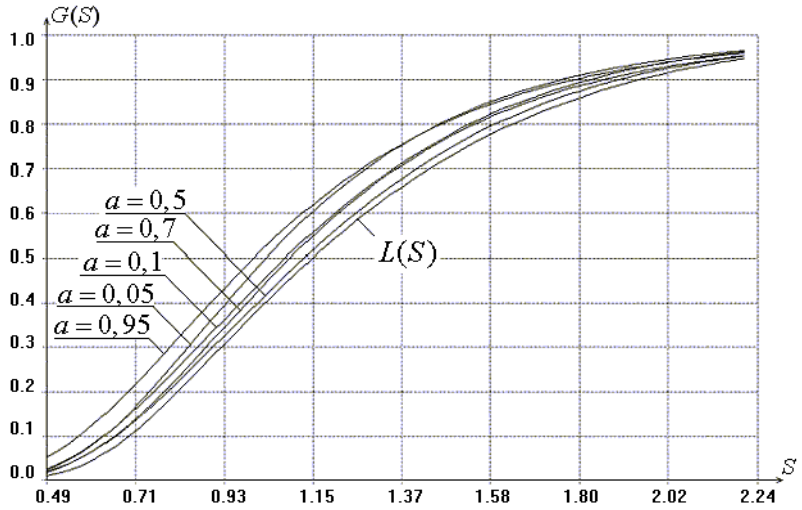

Рис. 2. Распределения статистики S_R^c при различной величине наблюдаемой области, $n = 100$, цензурирование слева

На рис. 3 представлены распределения статистики Колмогорова по цензурированным выборкам при проверке простой гипотезы о согласии с экспоненциальным законом. Здесь же приведены соответствующие предельные распределения статистики. Показано, что уже при потенциальном объеме выборки $n = 50$ в случае степени цензурирования $a < 0,5$ эмпирические распределения статистики Колмогорова хорошо согласуются с соответствующими предельными распределениями.

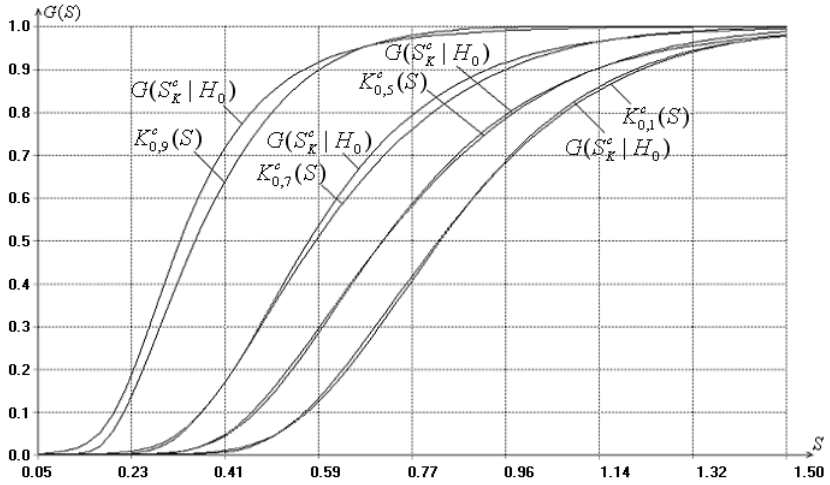

Рис. 3. Распределения статистики S_K^c при различной степени цензурирования и $n = 50$, цензурирование слева

В результате исследования распределений статистики Колмогорова найдены минимальные объемы выборок, при которых достигается хорошее согласие распределения статистики с соответствующим предельным законом для различных значений степени цензурирования a при проверке простых гипотез.

На практике чаще приходится иметь дело с проверкой сложных гипотез, когда в качестве параметров исследуемого распределения берут оценки параметров, полученные по тем же самым данным. Наиболее эффективным и универсальным по отношению к форме представления выборочных данных является метод максимального правдоподобия. Оценкой максимального правдоподобия неизвестного параметра по цензурированной слева и справа выборке является решение системы уравнений правдоподобия

$$n_1 \frac{\partial \ln P_1(\theta)}{\partial \theta_l} + \sum_{j=1}^{n_2} \frac{\partial \ln f(x_j, \theta)}{\partial \theta_l} + n_3 \frac{\partial \ln P_3(\theta)}{\partial \theta_l} = 0, \; l = \overline{1, m}, \tag{3}$$

где $P_1(\theta) = \int_{-\infty}^{x_{(1)}} f(x, \theta) dx$, $P_3(\theta) = \int_{x_{(2)}}^{\infty} f(x, \theta) dx$, n_i – количество наблюдений, попавших в i-ый интервал. В случае цензурирования только справа (только слева) в выражении исчезает первое (третье) слагаемое.

В результате исследования методами компьютерного моделирования распределений статистики критерия типа Колмогорова построены аппроксимации предельных распределений статистики для ряда законов, соответствующих проверяемой гипотезе H_0 (логарифмически нормального, экспоненциального, Вейбулла), при использовании для вычисления параметров этих законов ОМП по цензурированным данным. Как правило, распределения статистики в рассмотренных случаях хорошо описываются семейством логарифмически нормальных законов.

Заключение. Наиболее весомые результаты, полученные в классической математической статистике, имеют асимптотический характер. Однако на практике всегда имеют дело с ограниченными объемами наблюдений. В результате проведенных исследований показано, что при ограниченных объемах выборок распределения статистики Реньи существенно зависят от степени цензурирования. При высокой или, наоборот, при малой степени цензурирования предельным распределением $L(S)$ можно пользоваться лишь при полном объеме выборок $n > 1000$, что на практике редко выполнимо. Поэтому в практике статистического анализа цензурированных выборок не рекомендуется использовать критерий типа Реньи.

В то же время распределения статистики Колмогорова при проверке простых гипотез быстро сходятся к соответствующему предельному закону $K_a^c(S)$. Для различной величины степени цензурирования найдены объемы выборок, при которых обеспечивается корректное применение критерия Колмогорова.

Методами компьютерного моделирования проведено исследование распределений статистики типа Колмогорова при проверке сложных гипотез с использованием ОМП по цензурированным выборкам. Построены модели, аппроксимирующие предельные распределения статистики $G(S \mid H_0)$ при различных проверяемых гипотезах H_0.

1. Вероятность и математическая статистика: Энциклопедия. Под ред. Прохорова Ю.В. – М.: Большая Российская энциклопедия, 1999. – 910 с.
2. Мания Г.М. Статистическое оценивание распределений. - Тбилиси: Изд-во ТГУ, 1974. – 237 с.