Министерство образования и науки Российской Федерации

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Компьютерные технологии анализа данных и исследования статистических закономерностей: исследование мощности критериев проверки статистических гипотез

Исследование скорости сходимости распределений статистик критериев проверки статистических гипотез

Методические указания
к выполнению курсовых проектов
для студентов V-го курса ФПМИ
по направлению 010400.68
дневного отделения

Новосибирск, 2012

Методические указания предназначены для студентов, выполняющих курсовые проекты по курсу "Компьютерные технологии анализа данных и исследования статистических закономерностей" в третьем семестре (направление 010400.68 – Прикладная математика и информатика, магистерская программа – *Математическое и программное обеспечение информационных технологий моделирования и анализа данных*). Указания содержат необходимые сведения для выполнения курсового проекта, порядок выполнения, структуру оформления пояснительной записки и примерное содержание её разделов, варианты заданий.

Составители: доктор техн. наук, проф. *Б.Ю. Лемешко*,
канд. техн. наук, доц. *С.Н. Постовалов*,
канд. техн. наук, доц. *Е.В. Чимитова*

Работа подготовлена на кафедре
прикладной математики

Цель. *Изучение методики исследования скорости сходимости распределения статистики критерия к предельному с использованием компьютерных технологий.*

Методические указания

1. Постановка задачи

Пусть имеется выборка (выборки) наблюдений одномерной или многомерной случайной величины ξ: $X_1, X_2, \ldots, X_n$. О виде или свойствах случайной величины имеется некоторое предположение – гипотеза H_0.

Для проверки гипотезы H_0 сформулирован статистический критерий, который при заданной вероятности ошибки первого рода α определяет критическую область, при попадании в которую выборки гипотеза H_0 отвергается. Далее мы будем рассматривать только те критерии, у которых в явном виде задана одномерная статистика $S\left(X_1, X_2, \ldots, X_n\right)$, а критическая область представляет собой один или несколько интервалов значений статистики.

Пусть в случае верной гипотезы H_0 статистика критерия $S\left(X_1, X_2, \ldots, X_n\right)$ имеет функцию распределения $G_n(x)$, а при $n \to \infty$ - предельную функцию распределения $G(x)$.

Основной задачей курсового проекта является определение скорости сходимости $G_n(x)$ к $G(x)$, и определение объема выборки, при котором расстояние до предельного не превышает ε.

2. Определение скорости сходимости.

Пусть $\rho(G_n, G)$ - расстояние между двумя функциями $G_n(x)$ и $G(x)$. Например, свойствами расстояния обладает статистика Колмогорова:

$$D_n = \sup_{|x|<\infty} \left|G_n(x) - G(x)\right|.$$

Функцию $\rho(G_n, G)$ мы будем аппроксимировать степенной функцией вида an^{-b}. Будем говорить, что чем больше величина b, тем больше скорость сходимости распределения статистики к предельному закону.

3. Алгоритм моделирования закона распределения $G_n(x)$

Аналитическое нахождение функции распределения $G_n(x)$, как правило, представляет собой более сложную задачу, чем нахождение предельного закона распределения. Однако достаточно просто можно построить эмпирическую функцию распределения для $G_n(x)$, используя метод Монте-Карло.

Для этого нужно сгенерировать выборку значений статистик критерия объемом N: $\{S_1, S_2, \ldots, S_N\}$ и построить по ней эмпирическую функцию распределения $G_{n,N}(x)$:

1. Моделируется выборка наблюдений случайно величины ξ объемом n.
2. Вычисляется статистика критерия S.
3. Шаги 1-2 повторяются N раз. В результате получается выборка статистик $\{S_1, S_2, \ldots, S_N\}$.

4. Определение требуемого объема моделирования

Естественно, что эмпирическое распределение $G_{n,N}(x)$ отличается от $G_n(x)$, но величину отклонения δ (рис. 1) можно определить, используя центральную предельную теорему, согласно которой $P\left\{\left|G_{n,N}(x) - G_n(x)\right| < \delta\right\} \to 2\Phi(\delta) - 1 = \gamma,\ N \to \infty$. Отсюда можно определить объем моделирования N, при котором длина γ-доверительного интервала будет равна 2δ:

$$N = t_\gamma^2 \frac{G_n(x)(1 - G_n(x))}{\delta^2} \le \bar{N} = \frac{t_\gamma^2}{4\delta^2},\ t_\gamma = \Phi^{-1}\left(\frac{\gamma + 1}{2}\right).$$

Так, если задать δ=0,001, *а* γ=0,99, то $\bar{N} = 1\ 658\ 944$.

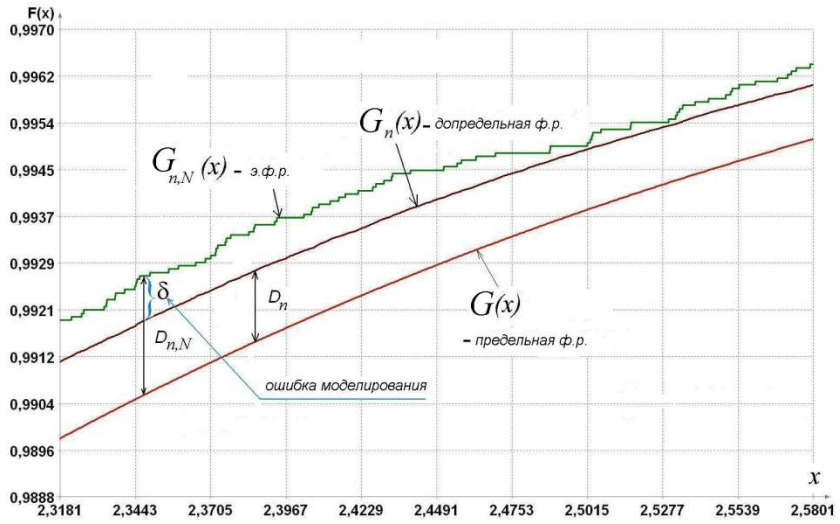

Рис. 1. Определение функции распределения $G_n(x)$

5. Аппроксимация расстояния до предельного закона степенной функцией

В результате моделирования должна получиться таблица расстояний следующего вида (таблица 1).

Таблица 1. Зависимость расстояния от n

n	$D_{n,N}$
5	0,029

6	0,023
7	0,018
8	0,015
9	0,013
…	…

Следует заметить, что увеличение n, когда $D_{n,N} < 2\delta$, уже не имеет смысла, т.к. в этом случае $D_{n,N}$ будет показывать ошибку моделирования, а не расстояние до предельного закона распределения.

Далее по таблице 1 можно подобрать функцию степенной регрессии an^{-b}, например, используя MS Excel.

6. Определение объема выборки, начиная с которого расстояние до предельного закона распределения не превышает заданного ε.

Используя найденное уравнение степенной регрессии, можно решить уравнение $an^{-b} = \varepsilon$ и найти объем выборки n, начиная с которого расстояние до предельного закона распределения не превышает заданного ε.

Порядок выполнения работы

1. Согласно варианту задания разработать программу для моделирования выборки статистик критерия.
2. Смоделировать выборки статистик для разных значений n.
3. Вычислить расстояние $D_{n,N}$ для каждого значения n.
4. Аппроксимировать зависимость расстояния до предельного закона распределения функцией an^{-b}.
5. Определить объем выборки, начиная с которого расстояние до предельного закона распределения не превышает 0,01.

Варианты заданий

№	Критерий	Источник	Закон распределения ξ	Дополнительные исследования	Уровень Сложности
Критерии согласия					
1	Колмогорова	[1]	Нормальное, Экспоненциальное, Коши	Статистика Колмогорова с поправкой Большева $\frac{6nD_n+1}{6\sqrt{n}}$	12
2	Смирнова	[1]	Логистическое, Вейбулла, Коши		10
3	Крамера-Мизеса-Смирнова	[1]	Лапласа, Экспоненциальное, Коши		10
4	Андерсона-Дарлинга	[1]	Нормальное, Рэлея, Коши		10
5	χ^2 Пирсона (простая гипотеза)	[1]	Нормальное, Вейбулла, Коши	Число интервалов группирования 2, 3, 5, 7, 10 АОГ, РВГ	15
6	χ^2 Рао-Робсона-Никулина (сложная гипотеза)	[1]	Нормальное, Экспоненциальное	Число интервалов группирования 2, 3, 5, 7, 10 АОГ, РВГ	20
7	Колмогорова для цензурированных справа или слева выборок (простая гипотеза)	[2]	Экспоненциальное, Рэлея	Степени цензурирования 5, 10, 20, 30, 40, 50, 60, 70, 80 Цензурирование I типа, II типа, слева, справа	20
8	Модифицированный медианный критерий	[3], с. 225	Нормальное, Экспоненциальное, Коши		12
9	Модифицированный критерий Колмогорова-Смирнова	[3], с. 225	Логистическое, Вейбулла, Коши		12
10	Модифицированный вероятностный критерий	[3], с. 226	Нормальное, Рэлея, Коши		12
Критерии экспоненциальности					
11	Большева	[3], [4]	Экспоненциальное		12
12	Гнеденко	[5] с. 1812	Экспоненциальное		12
13	Харриса	[5] с. 1813	Экспоненциальное		12
14	Холландера-Прошана	[5] с. 1814	Экспоненциальное		12
15	Гини	[5] с. 1815	Экспоненциальное		12
16	Эпштейна	[5] с. 1816	Экспоненциальное		12
17	Кокса-Оукса	[6] с. 33	Экспоненциальное		12
18	Эппса-Палли	[6] с. 36	Экспоненциальное		12
19	Ватсона	[3], с.282, с.222	Экспоненциальное		12
20	Купера	[3], с.282, с. 223	Экспоненциальное		12
Критерии равномерности					
21	Шермана	[3], с.319	Равномерный		12
22	Морана	[3], с.320	Равномерный		12
23	Ченга-Спиринга).	[3], с.322	Равномерный		12

24	Саркади-Косика).	[3], с.323	Равномерный		12
25	Хегази-Грина	[3], с.326	Равномерный		12
26	Гринвуда-Кэсенберри-Миллера	[3], с.320	Равномерный		12
Критерии однородности средних					
27	t-критерий Стьюдента при известных дисперсиях	[1]	Нормальное, Экспоненциальное, Коши	рассмотреть разные комбинации объемов выборок	15
28	t-критерий Стьюдента при неизвестных, но равных дисперсиях	[1]	Нормальное, Экспоненциальное, Коши	рассмотреть разные комбинации объемов выборок	15
29	t-критерий Стьюдента при неизвестных и неравных дисперсиях	[1]	Нормальное, Экспоненциальное, Коши	рассмотреть разные комбинации объемов выборок отличие дисперсий в 2х выборках в 2 раза, в 5 раз, в 10 раз	20
30	Уилкоксона	[1]	Нормальное, Экспоненциальное, Коши	рассмотреть разные комбинации объемов выборок рассмотреть также случай, когда дисперсии не равны (отличие в 2 раза, в 5 раз, в 10 раз)	20
31	Манна-Уитни	[1]	Нормальное, Экспоненциальное, Коши	рассмотреть разные комбинации объемов выборок рассмотреть также случай, когда дисперсии не равны (отличие в 2 раза, в 5 раз, в 10 раз)	20
Критерии однородности распределений					
32	Лемана-Розенблатта	[1]	Нормальное, Экспоненциальное, Коши	рассмотреть разные комбинации объемов выборок	15
33	Смирнова	[1]	Нормальное, Экспоненциальное, Коши	рассмотреть разные комбинации объемов выборок	
34	Катценбайссера-Хакля	[3] с. 228	Логистическое, Вейбулла, Коши	рассмотреть разные комбинации объемов выборок	15
Критерии однородности дисперсий (характеристик рассеяния)					
35	Левене	[1]	Нормальное, Лапласа, Вейбулла	выборочное среднее, выборочная медиана, усеченное среднее рассмотреть разные комбинации объемов выборок	15
36	Бартлетта	[1]	Нормальное, Макс. значений, Экспоненциальное	рассмотреть разные комбинации объемов выборок	15
37	Ансари-Бредли	[1]	Нормальное, Рэлея, Коши	рассмотреть разные комбинации объемов выборок	15
38	Муда	[1]	Нормальное, Максвелла, Коши	рассмотреть разные комбинации объемов выборок	15

39	Сижела-Тьюки	[1]	Лапласа, Экспоненциальное Коши,	рассмотреть разные комбинации объемов выборок	15
40	Кейпена	[1]	Логистическое, Экспоненциальное, Коши	рассмотреть разные комбинации объемов выборок	15
41	Клотца	[1]	Нормальное, Вейбулла, Коши	рассмотреть разные комбинации объемов выборок	15
Критерии выявления тренда					
42	Сериальный критерий Шведа-Эйзенхарта	[3], с.621			14
43	Критерий автокорреляции Кенуя	[3], с. 622			14
44	Критерий Блума-Кифера-Розенблатта	[3], с.623			14
45	Критерий Гёфдинга	[3], с.628			14

Литература

1. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход : [монография] / Б. Ю. Лемешко [и др.]. - Новосибирск, 2011. - 887 с. : ил., табл.
2. Лемешко Б.Ю., Чимитова Е.В., Плешкова Т.А. Проверка простых и сложных гипотез о согласии по цензурированным выборкам // Научный вестник НГТУ. - 2010. - № 4(41). – С.13-28. http://ami.nstu.ru/~headrd/seminar/publik_html/N_vestnik_2010.pdf
3. Кобзарь А.И. Прикладная математическая статистика для инженеров и научных работников. – М.: Физматлит, 2006. – 816 с.
4. Большев Л.Н. К вопросу о проверке «показательности». Вероятность и ее применения С. 542-544. (есть в электронном виде)
5. Ascher S. A survey of tests for exponentiality. Communications in Statistics - Theory and Methods, 1811-1825 (есть в электронном виде)
6. Henze N. and Meintanis S.G. Recent and classical tests for exponentiality: a partial review with comparisons. Metrika (2005) 61: 29–45 (есть в электронном виде)