Министерство образования и науки Российской Федерации

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Компьютерные технологии анализа данных и исследования статистических закономерностей

Методические указания
к выполнению лабораторных работ
для студентов I-го курса магистратуры ФПМИ (семестр 1)
по направлению 010400.68
дневного отделения

Новосибирск, 2012

Методические указания предназначены для студентов, выполняющих лабораторные работы по курсу "Компьютерные технологии анализа данных и исследования статистических закономерностей" (направление 010400.68). Указания содержат необходимые сведения для выполнения лабораторных работ, порядок выполнения, варианты заданий и контрольные вопросы, по которым осуществляется их защита.

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

Содержание

ВВЕДЕНИЕ

В лабораторных работах по компьютерным технологиям анализа данных и исследования статистических закономерностей применяется методика компьютерного моделирования фундаментальных статистических закономерностей. В основе методики лежит использование метода Монте-Карло для моделирования эмпирических распределений некоторых функций от случайных величин. Преимуществом метода Монте-Карло является несложность реализации процедур моделирования сложных статистических закономерностей, обычно не поддающихся определению аналитическими методами.

Выполнение работ носит исследовательский характер и в большинстве случаев опирается на разработанное программное обеспечение.

При выполнении лабораторных работ необходимо учитывать точность полученных результатов. Поскольку в большинстве работ требуется сравнивать эмпирические распределения, то малые объемы моделируемых выборок статистик или оценок могут нивелировать разницу в статистических методах, а иногда и давать некорректные (противоположные аналитическим выводам) результаты.

Согласие эмпирических распределений с теоретическими следует проверять, опираясь на статистические критерии, так как сравнение графиков распределений "на глаз" является субъективной процедурой и не учитывает возможной статистической погрешности, которая зависит от объемов выборок.

Отчет по лабораторной работе должен содержать цель, задание на выполнение, результаты исследований в соответствии с заданием, необходимые таблицы, графики, иллюстрирующие результаты, анализ результатов, общие выводы по работе.

Основной теоретический материал, необходимый для выполнения работ, изложен в учебном пособии [1], дополнительные материалы и программное обеспечение размещены на персональных сайтах авторов http://ami.nstu.ru/~headrd и http://postovalov.net.

Лабораторная работа № 1. Исследование свойств оценок параметров распределений вероятностей по эмпирическим данным

Цель работы. Вычисление оценок параметров распределений вероятностей по эмпирическим данным различными методами. Исследование асимптотических свойств оценок методом Монте-Карло.

Методические указания

1. Постановка задачи. Пусть в эксперименте наблюдается непрерывная случайная величина ξ с функцией распределения $F(x,\theta)$, где θ – вектор параметров. По выборке $\mathbf{X}_n = \{X_1,\ X_2,\ \ldots,\ X_n\}$ требуется оценить неизвестные параметры распределения.

2. Методы оценивания.

Метод максимального правдоподобия. Оценки максимального правдоподобия (ОМП) находятся в результате максимизации функции правдоподобия:

$$\hat{\theta}_{ОМП} = \arg\max_{\theta} \prod_{j=1}^{n} f(X_j, \theta).$$

Метод минимального расстояния. *MD*-оценки находятся в результате минимизации расстояния $\rho(F(x,\theta), F_n(x))$ между теоретической и эмпирической функциями распределения:

$$\hat{\theta}_{MD} = \arg\min_{\theta} \rho(F(x,\theta), F_n(x)),$$

где $F_n(x)$ – эмпирическая функция распределения. В качестве меры близости можно использовать следующие статистики:

а) Статистика Колмогорова:

$$D_n = \sup_x |F(x,\theta) - F_n(x)|$$

б) Статистика ω^2 Мизеса:

$$\omega_n^2 = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left\{F(X_{(i)},\theta) - \frac{2i-1}{2n}\right\}^2$$

в) Статистика Ω^2 Мизеса:

$$\Omega_n^2 = -1 - \frac{2}{n}\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(X_{(i)},\theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(X_{(i)},\theta))\right\}$$

Вычисление оценок по порядковым статистикам. Для нахождения оценок часто используются линейные комбинации порядковых статистик или выборочных квантилей. Такие оценки называются *L-оценками*. *L*-оценки обладают двумя важными для практического применения качествами: простотой вычислений и хорошими свойствами робастности.

При построении L-оценок по выборочным квантилям $z_1 < z_2 < .... < z_k$ рассматриваемого закона оценки находят в виде:

$$\hat{\theta} = \sum_{i=1}^{k} \alpha_i z(p_i), \quad z(p) = (X_{([np])} + X_{([np]+1)})/2,$$

где $X_{(i)}$ – i-я порядковая статистика, α_i и p_i – набор коэффициентов и вероятностей, которыми определяется конкретная оценка, n – объем выборки.

3. Свойства оценок.

Несмещенность. Оценка $\hat{\theta}$ называется *несмещенной*, если $M[\hat{\theta}(X_n)] = \theta$.

Состоятельность. Оценка $\hat{\theta}$ называется *состоятельной*, если $\hat{\theta}(X_n) \xrightarrow{p} \theta$, т.е. $\forall \varepsilon > 0 \ \lim\limits_{n\to\infty} P\left\{ \left| \hat{\theta}(X_n) - \theta \right| > \varepsilon \right\} = 0$.

Эффективность. Несмещенная оценка $\hat{\theta}$ называется *эффективной*, если $D[\hat{\theta}\ (X_n)] = J_n^{-1}(\theta)$, где $J(\theta) = M\left(\frac{\partial \ln L(\mathbf{X}_n, \theta)}{\partial \theta} \right)^2 = -M \frac{\partial^2 \ln L(\mathbf{X}_n, \theta)}{\partial \theta^2}$ – информационное количество Фишера.

Асимптотическая нормальность. Оценка $\hat{\theta}$ называется асимптотически нормальной, если ее распределение с ростом объема выборки стремится к нормальному закону распределения.

4. Исследование асимптотических свойств оценок методом Монте-Карло.

1. Моделируется N выборок по n наблюдений в каждой в соответствии с заданным законом распределения и фиксированными значениями параметров θ.
2. По каждой выборке вычисляются оценки параметров, в результате получается выборка $T_N = \{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N\}$.
3. Исследуется (идентифицируется) распределение случайной величины $\hat{\theta}(\mathbf{X}_n)$.
4. Пункты 1-3 повторяются при большем значении n.

По серии машинных экспериментов делаются выводы об изменении свойств оценок с ростом объема выборки.

Порядок выполнения

1. Определить количество экспериментов *N*, при котором с вероятностью 0.99 отклонение эмпирической функции распределения от теоретической не превышает величину 0.01.
2. Используя программу **ISW с**моделировать выборки, состоящие из оценок:

- максимального правдоподобия – $\hat{\theta}_{ОМП}(\mathbf{X}_n)$,

- минимального расстояния Колмогорова – $\hat{\theta}_K(\mathbf{X}_n)$,

- минимального расстояния ω^2 Мизеса – $\hat{\theta}_{\omega^2}(\mathbf{X}_n)$,

- минимального расстояния $\hat{\theta}_{\Omega^2}(\mathbf{X}_n)$ Мизеса,

- по порядковым статистикам (L-оценки)

при $n = 20, 50, 100, 1000$.

3. Построить графики эмпирических функций распределения смоделированных выборок, сравнить функции распределения оценок параметров при разных объемах и методах оценивания.

4. Исследовать полученные выборки на нормальность.

5. Заполнить таблицу (для каждого метода оценивания).

	n =20	*n =50*	*n =100*	*n =1000*
$\mathrm{M}[\hat{\theta}(\mathbf{X}_N)] - \theta$				
$P\left\{\left\|\hat{\theta}(\mathbf{X}_N) - \theta\right\| > 0.1\right\}$				
$\mathrm{D}[\hat{\theta}(\mathbf{X}_N)] - J_n^{-1}(\theta)$				

На основании таблицы сделать выводы о свойствах различных методов оценивания.

Варианты заданий

№ варианта	Распределение, параметр
1.	Экспоненциальное распределение, параметр масштаба.
2.	Распределение Лапласа, параметр масштаба.
3.	Нормальное распределение, параметр сдвига.
4.	Нормальное распределение, параметр масштаба.
5.	Распределение Коши, параметр сдвига.
6.	Распределение Коши, параметр масштаба.
7.	Логистическое распределение, параметр сдвига.
8.	Логистическое распределение, параметр масштаба.
9.	Распределение экстремальных значений (минимум), параметр сдвига.
10.	Распределение экстремальных значений (минимум), параметр масштаба.
11.	Распределение экстремальных значений (максимум), параметр сдвига.
12.	Распределение экстремальных значений (максимум), параметр масштаба.

Контрольные вопросы

1. Метод максимального правдоподобия.
2. Методы минимального расстояния.
3. Оценивание параметров по порядковым статистикам.

4. Свойства оценок по выборкам фиксированного объема.
5. Асимптотические свойства оценок.
6. Экспериментальное исследование свойств оценок.

Лабораторная работа № 2. Экспериментальное исследование робастности оценок

Цель работы. Исследование устойчивости оценок на наличие в выборке аномальных наблюдений. Исследование эффективности параметрической процедуры исключения аномальных наблюдений при использовании робастных оценок. Построение функций влияния Хампеля для ОМП. Исследование распределений статистик типа Граббса, предназначенных для анализа на аномальность сразу нескольких наблюдений.

Методические указания

1. Модель засорения выборки [3]. Пусть в эксперименте наблюдается непрерывная случайная величина ξ с функцией распределения

$$F_{\xi}(x) = (1-\nu)F(x,\theta) + \nu F_1(x,\theta_1), \qquad (2.1)$$

где ν – доля засорения выборки аномальными (с точки зрения закона $F(x,\theta)$) наблюдениями, подчиняющимися закону $F_1(x,\theta_1)$. По выборке $X_n = \{X_1,\ X_2,\ \ldots,\ X_n\}$ требуется оценить неизвестные параметры закона распределения $F(x,\theta)$.

2. Робастность. Под *робастностью* в статистике понимают нечувствительность к малым отклонениям от предположений (т.е. когда $\nu \to 0$).

3. Повышение робастности с помощью процедуры группирования. При группировании выборки теряется информация об индивидуальных наблюдениях, а фиксируется только количество наблюдений, попавших в интервалы группирования. В результате, небольшие отклонения от предполагаемого закона и аномальные выбросы не оказывают существенного влияния на оценки.

4. Параметрическая процедура отбраковки аномальных наблюдений состоит из двух этапов.

I этап. Одним из *робастных* методов находим оценки $\hat{\theta}$ параметров распределения $F(x,\hat{\theta})$.

II этап. Отбрасываем все наблюдения X_i: $X_i < \underline{d} \vee X_i > \overline{d}$. Пороговые значения определяются по формулам:

$$\underline{d} = F^{-1}(1-\sqrt[n]{1-\alpha},\hat{\theta}), \quad \overline{d} = F^{-1}(\sqrt[n]{1-\alpha},\hat{\theta}),$$

где n – объем выборки, α – уровень значимости критерия (вероятность ошибки первого рода, т.е. вероятность признать неаномальное наблюдение аномальным).

5. Функция влияния Хампеля. Для анализа робастности может использоваться функция влияния Хампеля:

$$IF(x;F,T)=\lim_{s\to 0}\frac{T((1-s)F+s\delta_x)-T(F)}{s},$$

где δ_x – единичная масса в точке x, $T(F)$ – статистика.

Если функция влияния неограничена, то резко выделяющиеся наблюдения могут приводить к существенным изменениям оценок или статистик. Чувствительность к большой ошибке характеризуется величиной:

$$\gamma^* = \sup_x \left| IF(x,F,T) \right|.$$

Функция влияния для асимптотически эффективных оценок, к которым относятся и ОМП, имеет вид:

$$IF(x,F,T)=J^{-1}(F(x,\theta))\frac{\partial \ln f(x,\theta)}{\partial\theta},$$

где $J^{-1}(F(x,\theta))=\int_{-\infty}^{\infty}\left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2 f(x,\theta)dx$ – информационное количество Фишера.

В случае группированных наблюдений для *k*-го интервала группирования (x_{k-1},x_k) функция влияния постоянна и равна

$$IF(x,F,T)=J^{-1}(F(x,\theta))\frac{\partial \ln P_k(\theta)}{\partial\theta},$$

где $P_k(\theta)=F(x_k,\theta)-F(x_{k-1},\theta)$.

6. Критерии типа Граббса. В случае принадлежности выборок нормальному закону для отбраковки аномальных наблюдений могут использоваться критерии типа Граббса. Наблюдаемая выборка упорядочивается (строится вариационный ряд) $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$.

Проверка на аномальность одного значения. При проверке на выброс $X_{(n)}$ статистика критерия Граббса имеет вид

$$G_n = (X_{(n)} - \overline{X})/S, \qquad (2.2)$$

где $\overline{X}=\frac{1}{n}\sum_{j=1}^{n}X_j$, $S^2=\frac{1}{n-1}\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2$.

При проверке на выброс наименьшего выборочного значения конкурирующая гипотеза H_1 предполагает, что $X_{(1)}$ принадлежит некоторому другому закону, существенно сдвинутому влево. В данном случае вычисляемая статистика принимает вид

$$G_1 = (\overline{X} - X_{(1)})/S. \qquad (2.3)$$

Максимальный или минимальный элемент выборки считается выбросом, если значение соответствующей статистики превысит критическое: $G_n \ge G_{n,1-\alpha}$ или $G_1 \ge G_{1,1-\alpha}$, где α – задаваемый уровень значимости.

Проверка на два выброса. В этом случае конкурирующая гипотеза H_1 может быть связана с предположением, что, например, некоторому другому закону принадлежат $X_{(n-1)}$ и $X_{(n)}$ (либо $X_{(1)}$ и $X_{(2)}$). При проверке на выброс одновременно двух наибольших значений статистика критерия Граббса имеет вид

$$G = S^2_{n-1,n} / S^2_0, \tag{2.4}$$

где

$$S^2_0 = \sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2, \quad S^2_{n-1,n} = \sum_{j=1}^{n-2} \left(X_j - \overline{X}_{n-1,n}\right)^2, \quad \overline{X}_{n-1,n} = \frac{1}{n-2} \sum_{j=1}^{n-2} X_j .$$

Для проверки на выброс одновременно двух наименьших величин $X_{(1)}$ и $X_{(2)}$ статистика критерия принимает вид

$$G = S^2_{1,2} / S^2_0, \tag{2.5}$$

где

$$S^2_{1,2} = \sum_{j=3}^{n} \left(X_j - \overline{X}_{1,2}\right)^2, \quad \overline{X}_{1,2} = \frac{1}{n-2} \sum_{j=3}^{n} X_j .$$

Оба значения ($X_{(n-1)}$, $X_{(n)}$ или $X_{(1)}$, $X_{(2)}$) считаются выбросами, если значение соответствующей статистики окажется ниже критического: $G < G_\alpha$.

Очевидно, как можно сформировать статистику критерия для проверки на наличие любой комбинации аномальных измерений.

Дополнительная информация размещена по адресу:
http://www.ami.nstu.ru/~headrd/seminar/publik_html/Z_lab_1.htm
http://www.ami.nstu.ru/~headrd/seminar/publik_html/NADEG_1.htm
http://www.ami.nstu.ru/~headrd/seminar/publik_html/Sibgim_1.htm
http://www.ami.nstu.ru/~headrd/seminar/publik_html/Z_lab_10.htm
http://www.ami.nstu.ru/~headrd/seminar/publik_html/Izm_T_6.htm

Порядок выполнения. Используя программу **ISW:**

1. Исследовать робастность оценок максимального правдоподобия (ОМП); ОМП по группированным данным; MD-оценок, минимизирующих расстояния, задаваемые статистиками Колмогорова, ω^2 и Ω^2 Мизеса; L-оптимальных оценок по выборочным квантилям. Для этого:
   - моделируются выборки с засорением (2.1), где параметр $\nu = 0, 0.1, 0.2$;
   - оцениваются параметры распределения $F(x, \theta)$ всеми методами;
   - сравниваются значения оценок при разных значениях ν.

2. Отбраковка аномальных наблюдений. Используя параметрическую процедуру отбраковки аномальных наблюдений очистить полученные в п.1 выборки от аномальных наблюдений и проверить, как изменились результаты применения критериев для проверки согласия эмпирического распределения с распределением $F(x,\theta)$ после удаления части наблюдений.

3. Построить функции влияния Хампеля для ОМП и ОМП по группированным данным.

4. Исследовать распределения статистик критериев типа Граббса, предназначенных для анализа на аномальность сразу нескольких наблюдений (*по заданию преподавателя*) в предположении о принадлежности выборки нормальному закону. Построить эмпирические распределения для статистик критериев типа Граббса, найти приближенные значения процентных точек. Для вариантов 1-5 применить критерий Граббса для отбраковки аномальных наблюдений по выборкам с засорением, полученным в п.1.

Варианты заданий.

№	$F(x,\theta)$	$F_1(x,\theta_1)$
1	Нормальное	Коши
2	Нормальное	Нормальное с масштабом 5
3	Нормальное	Нормальное со сдвигом 5
4	Нормальное	Нормальное с масштабом 10
5	Нормальное	Нормальное со сдвигом 10
6	Логистическое	Логистическое с масштабом 5
7	Логистическое	Логистическое со сдвигом 5
8	Логистическое	Логистическое с масштабом 10
9	Логистическое	Логистическое со сдвигом 10
10	Экспоненциальное	Экспоненциальное с масштабом 3
11	Экспоненциальное	Экспоненциальное с масштабом 5
12	Экспоненциальное	Экспоненциальное с масштабом 10

Контрольные вопросы.

1. Какие статистические процедуры называются робастными?
2. Как исследовать робастность процедуры оценивания параметров?
3. Как вычислить функцию влияния Хампеля?
4. Что представляет собой функция влияния Хампеля по группированным наблюдениям.
5. Как влияют объём выборки и уровень значимости α на параметрическую процедуру отбраковки аномальных наблюдений?
6. Можно ли применять критерии типа Граббса в случае нарушения предположений о нормальности исходной выборки? Если нет, то почему? Если да, то, в каком случае?

Лабораторная работа № 3. Экспериментальное исследование свойств критерия согласия χ^2 Пирсона.

Цель работы. Исследование влияния способов группирования на предельные распределения статистики критерия согласия χ^2 Пирсона при простых и сложных гипотезах (при использовании для вычисления оценок по негруппированным данным метода максимального правдоподобия).

Методические указания

В практике статистического анализа с необходимостью использования критериев согласия приходится сталкиваться при проверке простой гипотезы H_0: $f(x) = f(x, \theta_u)$, где $f(\cdot)$ - плотность распределения наблюдаемого закона, θ_u - известное истинное значение параметра (вектора параметров) закона, или при проверке сложной гипотезы H_0: $f(x) \in \{f(x) = f(x, \theta), \theta \in \Theta\}$, когда оценка $\hat{\theta}$ параметра предполагаемого закона распределения оценивается по этой же выборке.

Статистика критерия χ^2 Пирсона вычисляется в соответствии с соотношением

$$X_n^2 = n\sum_{i=1}^{k} \frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)}, \qquad (3.1)$$

где n_i – количество наблюдений, попавших в интервал, $n = \sum_{i=1}^{k} n_i$ $P_i(\theta)$ - вероятность попадания наблюдения в i-й интервал. При справедливости (простой) гипотезы H_0 предельным распределением $G(S|H_0)$ статистики является χ_r^2-распределение с числом степеней свободы $r = k - 1$. Если по выборке оценивалось p параметров закона в результате минимизации статистики X_n^2, статистика подчиняется χ_r^2-распределению с $r = k - p - 1$ степеней свободы. При справедливости конкурирующей гипотезы H_1 предельное распределение $G(S|H_1)$ представляет собой нецентральное χ_r^2-распределение с тем же числом степеней свободы и параметром нецентральности

$$\lambda = n\sum_{i=1}^{k} \frac{(P_i(\theta_1) - P_i(\theta))^2}{P_i(\theta)},$$

где $P_i(\theta_1)$ - вероятности попадания наблюдения в i-й интервал при конкурирующей гипотезе.

В случае проверки **сложных** гипотез и оценивании по выборке параметров распределений использование в качестве предельных χ^2_{k-p-1}-распределений справедливо лишь при определении оценок параметров по сгруппированным данным и использовании для оценивания статистики X_n^2:

$$\hat{\theta} = \arg\min_{\theta} X_n^2.$$

Это же справедливо в случае применения для сгруппированных данных метода максимального правдоподобия, оптимальных L-оценок по выборочным квантилям и некоторых других методов.

При использовании критериев согласия конкурирующая гипотеза H_1 (альтернатива) обычно не задается. Задавая конкретную альтернативу и имея возможность построить распределения статистик при истинности гипотезы H_0 ($G(S|H_0)$) и истинности гипотезы H_1 ($G(S|H_1)$), можно при заданном уровне значимости α (α - вероятность ошибки первого рода) вычислить мощность критерия $1-\beta$, которая определяет способность различения этих гипотез (β - вероятность ошибки второго рода).

Порядок выполнения

Исследовать распределения статистики критерия для простых и сложных гипотез (при использовании оценок максимального правдоподобия по негруппированным данным) при справедливой нулевой и конкурирующей гипотезах.

1. В соответствии с заданным наблюдаемым распределением (гипотеза H_0) смоделировать эмпирические распределения статистики критерия при простой гипотезе (а) (по выборке не оцениваются параметры), для сложной гипотезы (б) (по выборке оцениваются все параметры).
2. Идентифицировать построенные законы распределения (найти аналитические модели, наиболее хорошо описывающие эмпирические распределения).
3. Повторить п.1, моделируя выборку по закону, соответствующему гипотезе H_1, а оценивая по этой выборке параметры закона, соответствующего гипотезе H_0. *Для того, чтобы гипотеза H_1 была наиболее близка к гипотезе H_0, следует подобрать параметры распределения, соответствующего гипотезе H_1, из условия минимизации расстояния до распределения, соответствующего гипотезе H_1.*

Для моделирования распределения статистики при справедливой гипотезе H_0 ($G(S|H_0)$) следует генерировать псевдослучайные величины, соответствующие наблюдаемому закону, и оценивать его параметры (в случае сложной гипотезы). Для моделирования распределения этой же статистики при проверке той же самой гипотезы H_0, но при справедливой гипотезе H_1

($G(S|H_1)$), следует генерировать псевдослучайные величины по закону, соответствующему гипотезе H_1, а оценивать параметры закона, соответствующего гипотезе H_0.

4. В результате такого моделирования будут получены пары эмпирических распределений $G(S|H_0)$ и $G(S|H_1)$ для простой и сложной гипотез, по которым, задаваясь значением α, можно вычислить мощность критерия $1-\beta$ относительно конкурирующей гипотезы:

$$\int_{-\infty}^{S_\alpha} g(s|H_0)ds = 1-\alpha, \quad \int_{S_\alpha}^{\infty} g(s|H_1)ds = 1-\beta. \qquad (3.2)$$

Опираясь на эти распределения, построить оперативные характеристики критерия для проверки простой и сложной гипотез как функции вида $(1-\beta)(\alpha)$.

5. Повторяя пункты 3-4, исследовать влияние количества интервалов, способа группирования (равновероятное и асимптотически оптимальное) на мощность критерия.

Варианты заданий

№ п/п	Гипотеза H_0	Гипотеза H_1
1.	Нормальное	Логистическое
2.	Логистическое	Нормальное
3.	Нормальное	Лапласа
4.	Лапласа	Нормальное
5.	Логистическое	Лапласа
6.	Лапласа	Логистическое
7.	Гамма	Вейбулла-Гнеденко
8.	Вейбулла-Гнеденко	Гамма
9.	Экстремальных значений (минимум)	Экстремальных значений (максимум)
10.	Экстремальных значений (максимум)	Экстремальных значений (минимум)
11.	Нормальное	Коши
12.	Коши	Нормальное

Контрольные вопросы

1. Распределена статистика критерия χ^2 Пирсона при справедливости проверяемой гипотезы? При справедливости конкурирующей гипотезы?
2. Почему при асимптотически оптимальном группировании мощность критерия χ^2 Пирсона максимальна относительно близких конкурирующих гипотез?
3. Как идентифицировать закон распределения по заданной выборке?

Лабораторная работа № 4. Исследование распределения статистики и мощности критерия Рао-Робсона-Никулина

Цель работы: Исследовать влияние способов группирования на предельные распределения статистики критерия согласия типа χ^2 Никулина при простых и сложных гипотезах (при использовании для вычисления оценок метода максимального правдоподобия по негруппированным данным). Сравнение мощности критерия Никулина с мощностью критерия χ^2 Пирсона. Исследовать мощности критерия Никулина в зависимости от числа интервалов.

Методические указания

Никулиным М.С. [6] предложена модификация стандартной статистики (6.1), для которой предельным распределением является обычное распределение χ^2_{k-1} (количество степеней свободы не зависит от числа оцениваемых параметров). Неизвестные параметры распределения $F(x,\theta)$ в этом случае должны оцениваться по негруппированным данным методом максимального правдоподобия. При этом вектор вероятностей попадания в интервалы $p=(p_1,\ldots,p_k)^{\mathrm{T}}$ предполагается заданным и граничные точки интервалов определяются соотношениями $x_i(\theta)=F^{-1}(p_1+\ldots+p_i)$, $i=\overline{1,(k-1)}$.

Предложенная статистика отличается от X_n^2 только при проверке сложных гипотез и имеет вид

$$Y_n^2(\theta)=X_n^2+n^{-1}a^{\mathrm{T}}(\theta)\Lambda(\theta)a(\theta),$$

где X_n^2 вычисляется в соответствии с (3.1). Элементы и размерность матрицы

$$\Lambda(\theta)=\left[J(\theta_l,\theta_j)-\sum_{i=1}^{k}\frac{w_{\theta_l i}w_{\theta_j i}}{p_i}\right]^{-1}_{m\times m}$$

определяются оцениваемыми компонентами вектора параметров θ, $J(\theta_l,\theta_j)$ - элементы информационной матрицы $\mathbf{J}(\theta)$, $a(\theta_l)=w_{\theta_l 1}n_1/p_1+\ldots+w_{\theta_l k}n_k/p_k$ - элементы вектора $a(\theta)$, величины $w_{\theta_l i}$ определяются соотношением

$$w_{\theta_l i}=-f\left[x_i(\theta),\theta\right]\frac{\partial x_i(\theta)}{\partial\theta_l}+f\left[x_{i-1}(\theta),\theta\right]\frac{\partial x_{i-1}(\theta)}{\partial\theta_l}.$$

Порядок выполнения.

1. Проверить, насколько хорошо распределение статистики Y_n^2 при справедливой нулевой гипотезе $G(Y_n^2|H_0)$ соответствует распределению χ^2_{k-1}
2. Исследовать влияние типа группирования на распределение статистики Никулина.

3. Построить оперативные характеристики критерия для простой и сложных гипотез как функции вида $(1-\beta)(\alpha)$. Сравнить мощность предложенного Никулиным критерия с мощностью критерия χ^2 Пирсона (результаты по мощности критерия χ^2 Пирсона взять из отчета по лабораторной работе № 3). *Для того чтобы распределение, соответствующее гипотезе H_1 было наиболее близким к распределению, соответствующему гипотезе H_0, следует подобрать параметры распределения, соответствующего гипотезе H_1, из условия минимизации расстояния до распределения, соответствующего основной гипотезе.*
4. Исследовать мощность критерия Никулина в зависимости от числа интервалов и выполнить сравнение с аналогичными исследованиями по критерию χ^2 Пирсона.

Варианты заданий совпадают с заданиями из лабораторной работы № 3

Контрольные вопросы

1. Сравнить критерии χ^2 Пирсона и Рао-Робсона-Никулина. Когда нельзя применять критерий Рао-Робсона-Никулина?
2. Вычислить поправку Никулина для экспоненциального распределения с оцениванием параметра масштаба.
3. Вычислить поправку Никулина для нормального распределения с с оцениванием параметра сдвига.
4. Вычислить поправку Никулина для нормального распределения с оцениванием параметра масштаба.

Лабораторная работа № 5. Экспериментальное исследование предельных распределений статистик непараметрических критериев согласия.

Цель работы. Исследование распределений статистик непараметрических критериев согласия при проверке простых и различных сложных гипотез.

Методические указания

В практике статистического анализа с необходимостью использования критериев согласия приходится сталкиваться как при проверке простой гипотезы H_0: $F(x)=F(x,\theta_{\grave{e}})$, где $F(\cdot)$ - плотность распределения наблюдаемого закона, θ_u - известное истинное значение параметра (вектора параметров) закона, так и при проверке сложной гипотезы H_0: $F(x)\in\{F(x,\theta),\ \theta\in\Theta\}$. Проблемы имеются, когда оценка $\hat{\theta}$ параметра предполагаемого закона распределения вычисляется по той же самой выборке, по которой проверяется согласие. Если оценка $\hat{\theta}$ вычисляется по другой выборке, то ситуация не

отличается от проверки простой гипотезы. В дальнейшем, если оценка параметра $\hat{\theta}$ вычисляется по этой же выборке, будем обозначать сложную гипотезу $H_0: F(x) = F(x, \hat{\theta})$.

В случае проверки сложных гипотез предельные распределения статистик непараметрических критериев согласия типа Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса, при справедливости нулевой гипотезы $H_0: F(x) = F(x, \theta)$ отличаются от предельных распределений классических статистик (когда по выборке не оцениваются параметры) [7]. В случае сложной гипотезы предельные распределения статистик зависят: от вида наблюдаемого закона, от количества и типа оцениваемых параметров этого распределения, от используемого метода оценивания параметров. А при ограниченных объемах выборок распределения статистик существенно зависят и от объема выборки.

Знание распределения статистики при проверке одной и той же гипотезы, но при различных истинных гипотезах (H_0 или H_1) позволяет определить мощность критерия, т.е. его способность различать эти гипотезы. Задавая конкретную альтернативу и имея возможность построить распределения статистик при истинности нулевой гипотезы H_0 ($G(S|H_0)$) и истинности альтернативы H_1 ($G(S|H_1)$), можно при заданном уровне значимости α вычислить мощность критерия $1-\beta$, которая определяет способность различения этих гипотез.

Для построения распределения $G(S|H_0)$ (распределения статистики при справедливой гипотезе H_0) следует моделировать псевдослучайные величины, соответствующие наблюдаемому закону и оценивать его параметры, после чего вычислять значение требуемой статистики S. А для построения распределения $G(S|H_1)$ (распределения той же статистики при проверке той же самой сложной гипотезы H_0, но при справедливой гипотезе H_1) следует моделировать псевдослучайные величины по закону, соответствующему гипотезе H_1, а оценивать параметры закона, соответствующего гипотезе H_0.

Порядок выполнения

1. Смоделировать распределение статистики S для заданного критерия согласия при простой гипотезе H_0. Сравнить полученное эмпирическое распределение с предельным распределением классической статистики.
2. Смоделировать распределение статистики S для этого же критерия согласия при сложной гипотезе H_0. Попытаться идентифицировать полученное эмпирическое распределение, используя систему статистического анализа **ISW**.
3. Смоделировать распределения статистики S исследуемого критерия согласия при проверке простой и сложной гипотез H_0 при справедливой

гипотезе H_1. *Для того чтобы распределение, соответствующее гипотезе* H_1 *было наиболее близким к распределению, соответствующему гипотезе* H_0, *следует подобрать параметры распределения, соответствующего гипотезе* H_1, *из условия минимизации расстояния до распределения, соответствующего основной гипотезе.*

4. Построить оперативные характеристики критерия для простой и сложных гипотез как функции вида $(1-\beta)(\alpha)$. Сравнить мощность непараметрических критериев с мощностью критериев χ^2 Пирсона и Рао-Робсона-Никулина (оперативные характеристики критериев χ^2 Пирсона и Рао-Робсона-Никулина взять из отчетов по лабораторным работам № 3-4)

Варианты заданий совпадают с заданиями из лабораторной работы № 3.

Контрольные вопросы

1. Критерий Колмогорова при проверке простых и сложных гипотез?
2. Критерий Крамера-Мизеса-Смирнова при проверке простых и сложных гипотез?
3. Критерий Андерсона-Дарлинга при проверке простых и сложных гипотез?
4. Влияет ли метод оценивания параметров на распределения статистик непараметрических критериев согласия при проверке сложной гипотезы?
5. Влияют ли значения параметров на распределения статистик непараметрических критериев согласия при проверке сложной гипотезы?

Лабораторная работа № 6. Исследование критериев проверки отклонения от нормального закона

Цель работы. Исследование распределений статистик критериев, используемых при проверке отклонения эмпирических распределений наблюдаемых величин от нормального закона (в том числе критериев проверки гипотез о симметричности и о значении эксцесса при различных наблюдаемых законах). Исследование распределений статистик критериев Шапиро-Уилка, Эппса-Палли, Д'Агостино. Исследование и сравнение мощности критериев относительно близких конкурирующих гипотез.

Методические указания

1. **Критерий симметричности** предназначен для проверки гипотез о симметричности наблюдаемого закона (против наличия асимметрии) при объемах выборки $8 \le n \le 5000$. Статистика критерия имеет вид

$$\sqrt{\hat{\beta}_1} = \frac{\hat{\mu}_3}{\hat{\sigma}^3}. \qquad (6.1)$$

Проверяется гипотеза H_0: $\sqrt{\beta_1}$ =0 против альтернативы $\sqrt{\beta_1}$ >0 (положительная асимметрия) или $\sqrt{\beta_1}$ <0 (отрицательная асимметрия).

2. **Критерий проверки на эксцесс** рассматривается при объемах выборок $8 \le n \le 5000$. Статистика критерия проверки на значение эксцесса имеет вид

$$\beta_2 = \frac{\hat{\mu}_4}{\hat{\sigma}^4}. \qquad (6.2)$$

Проверяется гипотеза вида H_0: β_2=3 против альтернативы β_2>3 (больший эксцесс) или β_2<3 (меньший эксцесс).

3. В **критерии Шапиро-Уилка** для вариационного ряда $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$, полученного по наблюдаемой выборке $X_1, X_2, ..., X_n$, вычисляют величину

$$S = \sum_k a_k \left[X_{(n+1-k)} - X_{(k)} \right],$$

где индекс k изменяется от 1 до $n/2$ или от 1 до $(n-1)/2$ при четном и нечетном n соответственно. Коэффициенты a_k приведены в ГОСТ Р ИСО 5479-2002 и первоисточниках. Статистика критерия имеет вид

$$W = S^2 \Big/ \sum_{i-1}^{n} \left(X_i - \overline{X} \right)^2. \qquad (6.3)$$

Гипотеза о нормальности отвергается при малых значениях статистики W.

4. Статистика **критерия Эппса-Палли** для наблюдаемой выборки $X_1, X_2, ..., X_n$ имеет вид

$$T_{EP} = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n} \sum_{k=2}^{n} \sum_{j=1}^{k-1} \exp\left\{ -\frac{\left(X_j - X_k \right)^2}{2\hat{\mu}_2} \right\} - \sqrt{2} \sum_{j=1}^{n} \exp\left\{ -\frac{\left(X_j - \overline{X} \right)^2}{4\hat{\mu}_2} \right\}, \qquad (6.4)$$

где $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$. Выборка может быть неупорядочена, порядок наблюдений произволен, но он должен быть неизменным в течение всех проводимых вычислений. Гипотезу о нормальности отвергают при больших значениях статистики.

5. **Модификация D'Agostino** критерия **проверки на симметричность**. В данной модификации на основании следующих соотношений статистика (6.1) преобразуется в статистику z_1, приближенно подчиняющуюся стандартному нормальному закону:

$$b=\frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)},\quad \omega^2=-1+\{2(b-1)\}^{1/2},$$

$$\delta=\frac{1}{\left\{\log(\sqrt{\omega^2})\right\}^{1/2}},\qquad y=\sqrt{\beta_1}\left\{\frac{\omega^2-1}{2}\cdot\frac{(n+1)(n+3)}{6(n-2)}\right\}^{1/2},$$

$$z_1=\delta\log\left\{y+(y^2+1)^{1/2}\right\}. \qquad (6.5).$$

6. **Модификация D'Agostino** критерия одновременной **проверки на симметричность и значение эксцесса**. Здесь предложено преобразование статистик (6.2) и (6.1) к статистике z_2, приближенно распределенной в соответствии со стандартным нормальным законом. Преобразование осуществляется с помощью следующих соотношений:

$$\delta=(n-3)(n+1)(n^2+15n-4),\quad a=\frac{(n-2)(n+5)(n+7)(n^2+27n-70)}{6\delta}$$

$$c=\frac{(n-7)(n+5)(n+7)(n^2+2n-5)}{6\delta},\ k=\frac{(n+5)(n+7)(n^3+37n^2+11n-313)}{12\delta},$$

$$\alpha=a+\beta_1 c,\qquad \chi=(\beta_2-1-\beta_1)2k,$$

$$z_2=\left\{\left(\frac{\chi}{2\alpha}\right)^{1/3}-1+\frac{1}{9\alpha}\right\}(9\alpha)^{1/2}\ . \qquad (6.6)$$

Дополнительная информация по критериям доступна по адресам:
http://www.ami.nstu.ru/~headrd/seminar/publik_html/Izm_T_7.htm
http://www.ami.nstu.ru/~headrd/seminar/Kontrol_Q

Порядок выполнения

Используя программу **ISW**, исследовать описанные выше критерии проверки отклонения наблюдаемых данных от нормального закона. В частности:

1. Исследовать зависимость распределений статистик (6.1) и (6.2) от объема выборок в случае принадлежности наблюдений нормальному закону. При некотором объеме выборок смоделировать эмпирические распределения статистик (6.1) и (6.2) при законе, отличающемся от нормального. Сравнить распределения статистик со случаем нормальности наблюдаемого закона.
2. Исследовать распределения статистик критерия Шапиро-Уилка при различных объемах выборок. Оценить близость получаемых эмпирических распределений статистик к «теоретическим» по процентным точкам таблиц, соответствующим данному критерию. При некотором объеме выборок (n=10) смоделировать распределение статистики критерия при семействе экспоненциальных распределений (двустороннее экспоненциальное) при параметре формы, равном 4÷6. Сравнить с ситуацией, соответствующей

справедливой проверяемой гипотезе о нормальном законе. Оценить мощность критерия.

3. Аналогично предыдущему пункту исследовать распределения статистики критерия Эппса-Палли.
4. Исследовать распределения статистики z_1. Проверить близость эмпирических распределений статистики стандартному нормальному закону.
5. Исследовать распределения статистики z_2. Проверить близость эмпирических распределений статистики стандартному нормальному закону.
6. Оценить мощность критериев со статистиками (6.3), (6.4), (6.5), (6.6) относительно заданной альтернативы.

Варианты заданий

№	$F_1(x,\boldsymbol{\theta}_1)$	№	$F_1(x,\boldsymbol{\theta}_1)$
1	Двустороннее экспоненциальное с параметром формы 3	8	Лапласа
2	Двустороннее экспоненциальное с параметром формы 5	9	Коши
3	Двустороннее экспоненциальное с параметром формы 10	10	Двустороннее экспоненциальное с параметром формы 0.5
4	Двустороннее экспоненциальное с параметром формы 7	11	Гамма-распределение с параметром формы 5
5	Распределение экстремальных значений (минимум)	12	Распределение экстремальных значений (максимум)
6	Экспоненциальное	13	Нормальное с масштабом 10
7	Логистическое	14	Нормальное со сдвигом 10

Контрольные вопросы

1. Какие альтернативы плохо различимы критериями проверки отклонения от нормальности Шапиро-Уилка и Эппса-Палли?
2. Можно ли по результатам проверки симметричности закона и проверки эксцесса утверждать о нормальности проверяемых данных?
3. Начиная с каких объемов выборок, критерии согласия не уступают по мощности критериям проверки на нормальность (Шапиро-Уилка, Эппса-Палли)?

Литература

1. Лемешко Б.Ю., Постовалов С.Н. Компьютерные технологии анализа данных и исследования статистических закономерностей: Учебное пособие. – Новосибирск: Изд-во НГТУ, 2004. – 119 с.
2. Ермаков С.М., Михайлов Г.А. Статистическое моделирование. – Москва: Наука, 1982. – 296 с.
3. Хьюбер П. Робастность в статистике. М.: Мир, 1984. – 303 с.
4. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983. - 416 с.
5. Денисов В.И., Лемешко Б.Ю., Постовалов С.Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть I. Критерии типа χ^2. - Новосибирск: Изд-во НГТУ, 1998. - 126 с.
6. Мирвалиев М., Никулин М.С. Критерии согласия типа хи-квадрат / Заводская лаборатория. 1992. Т. 58. № 3. С.52-58.
7. Лемешко Б.Ю., Постовалов С.Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть II. Непараметрические критерии. - Новосибирск: Изд-во НГТУ, 1999. - 85 с.

Приложение. Законы распределения наблюдаемых случайных величин

Закон распределения, параметры, область определения	Функция плотности $f(x,\theta)$
Экспоненциальный, $\theta_0 > 0$, $x \in [0,+\infty]$	$\frac{1}{\theta_0} e^{-x/\theta_0}$
Полунормальный, $\theta_0 > 0$, $x \in [0,+\infty]$	$\frac{2}{\theta_0\sqrt{2\pi}} e^{-x^2/2\theta_0^2}$
Рэлея, $\theta_0 > 0$, $x \in [0,+\infty]$	$\frac{x}{\theta_0^2} e^{-x^2/2\theta_0^2}$
Максвелла, $\theta_0 > 0$, $x \in [0,+\infty]$	$\frac{2x^2}{\theta_0^3\sqrt{2\pi}} e^{-x^2/2\theta_0^2}$
Лапласа, $\theta_1 \in R$, $\theta_0 > 0$,	$\frac{1}{2\theta_0} e^{-\lvert x-\theta_1\rvert/\theta_0}$

$x\in[-\infty,+\infty]$	
Нормальный, $\theta_1\in R$, $\theta_0>0$, $x\in[-\infty,+\infty]$	$\frac{1}{\theta_0\sqrt{2\pi}}e^{-\frac{(x-\theta_1)^2}{2\theta_0^2}}$
Логнормальный, $\theta_1\in R$, $\theta_0>0$, $x\in[0,+\infty]$	$\frac{1}{x\theta_0\sqrt{2\pi}}e^{-(\ln x-\theta_1)^2/2\theta_0^2}$
Коши, $\theta_1\in R$, $\theta_0>0$, $x\in[-\infty,+\infty]$	$\frac{\theta_0}{\pi[\theta_0^2+(x-\theta_1)^2]}$
Логистический, $\theta_0\in R$, $\theta_1>0$, $x\in[-\infty,+\infty]$	$\frac{\pi}{\theta_0\sqrt{3}}\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\Big/\left[1+\exp\left\{-\frac{\pi(x-\theta_1)}{\theta_0\sqrt{3}}\right\}\right]^2$
Экстремальных значений (максимум), $\theta_0>0$, $\theta_1\in R$, $x\in[-\infty,+\infty]$	$\frac{1}{\theta_0}\exp\left\{-\frac{x-\theta_1}{\theta_0}-\exp\left(-\frac{x-\theta_1}{\theta_0}\right)\right\}$
Экстремальных значений (минимум), $\theta_0>0$, $\theta_1\in R$, $x\in[-\infty,+\infty]$	$\frac{1}{\theta_0}\exp\left\{\frac{x-\theta_1}{\theta_0}-\exp\left(\frac{x-\theta_1}{\theta_0}\right)\right\}$
Вейбулла-Гнеденко, $\theta_0>0$, $\theta_1>0$, $x\in[0,+\infty]$	$\frac{\theta_0 x^{\theta_0-1}}{\theta_1^{\theta_0}}\exp\left\{-\left(\frac{x}{\theta_1}\right)^{\theta_0}\right\}$
Гамма, $\theta_0>0$, $\theta_1>0$, $\theta_2\in R$, $x\in[\theta_2,+\infty]$	$\frac{1}{\theta_1^{\theta_0}\Gamma(\theta_0)}(x-\theta_2)^{\theta_0-1}e^{-(x-\theta_2)/\theta_1}$
Бета-распределение I-го рода, $\theta_0>0$, $\theta_1>0$, $\theta_2>0$, $x\in[0,\theta_2]$	$f(x)=\frac{\Gamma(\theta_0+\theta_1)}{\lambda^{\theta_0+\theta_1-1}\Gamma(\theta_0)\Gamma(\theta_1)}x^{\theta_0-1}(\theta_2-x)^{\theta_1-1}$
Бета-распределение II - го рода, $\theta_0>0$, $\theta_2>0$, $\theta_3>0$, $x\in[\theta_1,+\infty]$	$\frac{1}{\theta_0\cdot\mathrm{B}(\theta_2,\theta_3)}\left(\frac{x-\theta_1}{\theta_0}\right)^{\theta_2-1}\left(1+\frac{x-\theta_1}{\theta_0}\right)^{-\theta_2-\theta_3}$

Sb-Джонсона, $\theta_1 \in R$, $\theta_3 \in R$, $\theta_1 > 0$, $\theta_2 > 0$, $x \in [\theta_3, \theta_3 + \theta_2]$	$\frac{\theta_1 \theta_2}{(x-\theta_3)(\theta_2+\theta_3-x)} \exp\left\{-\frac{1}{2}\left[\theta_0 - \theta_1 \ln \frac{x-\theta_3}{\theta_2+\theta_3-x}\right]^2\right\}$
Двустороннее экспоненциальное, $\theta_1 \in R$, $\theta_0 > 0$, $\theta_2 > 0$, $x \in [-\infty, +\infty]$	$f(x) = \frac{\theta_2}{2\theta_1 \Gamma(1/\theta_2)} e^{-\left\lvert \frac{x-\theta_1}{\theta_0} \right\rvert^{\theta_2}}$