Министерство образования и науки Российской Федерации
Новосибирский государственный технический университет
Кафедра теоретической и прикладной информатики

Курсовой проект

на тему

«Исследование скорости сходимости распределения статистики к предельному закону критерия нормальности Жарка-Бера»

по курсу

«Компьютерные технологии анализа данных и исследования статистических закономерностей»

Факультет: ФПМИ
Группа: ПММ-61
Студент: Горбунов К. К.
Преподаватель: проф. Постовалов С. Н.

Новосибирск, 2017 г.

Содержание

Введение **3**

1 Постановка задачи **4**

1.1 Определение скорости сходимости . . . 4

1.2 Алгоритм построения закона распределения $G_n(x)$ . . . 5

1.3 Определение требуемого объема моделирования . . . 5

1.4 Аппроксимация расстояния до предельного закона степенной функцией . . . 6

1.5 Особенности определения скорости сходимости . . . 6

1.6 Определение объема выборки, начиная с которого расстояние до предельного закона распределения не превышает заданного ε . . . 7

1.7 Проверка полученного результата . . . 7

1.8 Порядок выполнения курсового проекта . . . 8

2 Краткие теоретические сведения **9**

2.1 Статистика критерия Жарка-Бера . . . 9

3 Результаты исследований **10**

3.1 Выбор параметров моделирования . . . 10

3.2 Моделирование распределения статистики . . . 10

3.3 Аппроксимация расстояния до предельного закона степенной функцией . . . 11

3.4 Определение объема выборки n, начиная с которого расстояние до предельного закона распределения не превышает заданного ε . . . 12

3.5 Контрольльный эксперимент . . . 12

4 Заключение **13**

Приложение А Исходные тексты программ **14**

A.1 Фрагменты исходных текстов 14

Введение

Целью работы является изучение методики исследования скорости сходимости распределения статистики критерия к предельному распределению с использованием компьютерных технологий.

Принадлежность наблюдаемых данных нормальному закону является необходимой предпосылкой для корректного применения большинства классических методов математической статистики, используемых в задачах обработки измерений, стандартизации и контроля качества. Поэтому проверка на отклонение от нормального закона является частой процедурой в ходе проведения измерений, контроля и испытаний, имеющей особое значение, так как далеко не всегда ошибки измерений, связанные с приборами, построенными на различных физических принципах, или ошибки наблюдений некоторого контролируемого показателя подчиняются нормальному закону. В таких случаях применение классического аппарата, опирающегося на предположение о нормальности наблюдаемого закона, оказывается некорректным и может приводить к неверным выводам [1].

В данной работе рассматривается совместный критерий проверки на симметричность и эксцесс (Jarque-Bera, Жарка-Бера)[2] — критерий проверки принадлежности выборки нормальному закону распределения. В связи с этим и сказанным в абзаце выше предполагается считать тему работы актуальной.

1. Постановка задачи

Пусть имеется выборка (выборки) наблюдений одномерной или многомерной случайной величины $\xi : X_1, X_2, \ldots, X_n$. О виде или свойствах случайной величины имеется некоторое предположение — гипотеза H_0. Для проверки гипотезы H_0 сформурирован статистический критерий, который при заданной вероятности ошибки первого года α определяет критическую область, при попадании в которую выборки гипотеза H_0 отвергается. Рассматриваются только те критерии, у которых в явном виде задана одномерная статистика $S(X_1, X_2, \ldots, X_n)$, а критическая область представляет собой один или несколько интервалов значений статистики.

Пусть в случае верной гипотезы H_0 статистика критерия $S(X_1, X_2, \ldots, X_n)$ 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.

1.1 Определение скорости сходимости

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

$$D_n = \sup_{|x|<\infty} |G_n(x) - G(x)| . \tag{1}$$

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

1.2 Алгоритм построения 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 статистических испытаний 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 случайной величины ξ объемом n.
2. Вычисляется статистика критерия S.
3. Шаги 1–2 повторяются N раз. В результатие получается выборка статистик $\{S_1, S_2, \ldots, S_N\}$.

1.3 Определение требуемого объема моделирования

Естественно, что эмпирическое распределение $G_{n,N}(x)$ отличается от $G_n(x)$, но величину отклонения ε_N можно определить по теореме Колмогорова и подобрать такое N, при котором величина погрешности моделирования будет меньше, чем расстояние $\rho(G_n, G)$. Если ошибка моделирования будем больше, чем расстояние до предельного закона, то тогда восстановить зависимость расстояния от объема выборки будет очень сложно.

По теореме Колмогорова

$$\lim_{n\to\infty} P\left\{\sqrt{N} \sup_{|x|<\infty} |G_{n,N}(x) - G_n(x)| < t\right\} = K(t), \tag{2}$$

где $K(t)$ — функция Колмогорова. Отсюда можно найти такое N, при котором величина погрешности моделирования $\varepsilon_N = \sup\limits_{|x|<\infty} |G_{n,N}(x) - G_n(x)|$ не превосходит заданного значения с некоторой доверительной вероятностью. Так, например, если мы хотим, чтобы погрешность моделирования ε_N не превышала 0.001 с вероятностью 0.99, то мы должны взять объем выборки статистик, равный

$$N = \left[\left(\frac{K^{-1}(0.99)}{0.001} \right)^2 \right] + 1 \approx \left[\left(\frac{1.62762}{0.001} \right)^2 \right] + 1 = 2649147. \tag{3}$$

Отметим, что этот объем больше, чем если бы мы рассматривали отклонение в одной точке, используя центральную предельную теорему:

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

Так, если задать $\delta = 0.001$, а $\gamma = 0.99$, то $\overline{N} = 1658944$.

1.4 Аппроксимация расстояния до предельного закона степенной функцией

В результате моделирования должна получиться таблица расстояний со строками вида $(n_i, D_{n_i,N})$, $i = 1, 2, \ldots, q$, где q — число строк таблицы.

Начальное значение n и шаг его изменения могут быть большими, причем шаг увеличения n может быть неравномерным. Следует отметить, 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 данным из таблицы можно подобрать функцию степенной регрессии an^{-b}.

1.5 Особенности определения скорости сходимости

Качество аппроксимации степенной регрессией зависимости расстояния между распределением статистики и предельным законом можно определить по коэффициенту детерминации. При хорошей аппроксимации коэффициент детерминации должен быть близок к 1, а точки на графике должны быть случайно разбросаны по обе стороны от линии тренда.

Если аппроксимация плохая, то можно использовать следующие рекомендации по её улучшению:

1. Увеличить объем моделирования N до рекомендуемого значения (3). Недостаточный объем моделирования приводит к большей ошибке в оценивании расстояния между распределением статистики и предельным законом.

2. Отсечь в таблице строки с маленьким значением n. Так как скорость сходимости является асимптотической величиной, то отбрасыванием маленьких объемов n можно существенно улучшить аппроксимацию.

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

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

1.7 Проверка полученного результата

Естественно, что, оценивая скорость сходимости, мы можем допустить ошибку, и поэтому, требуется провести контрольный эксперимент, чтобы убедиться в надежности наших выводов.

Контрольный эксперимент заключается в следующем. Для найденного значения n^* в п.1.6 моделируется выборка статистик объемом $N = 2649147$ и вычисляется расстояние до предельного закона $D_{n,N}$. В случае корректного определения скорости сходимости величина $D_{n,N}$ должна отклоняться от 0.01 не более чем на 0.001 с доверительной вероятностью 0.99.

1.8 Порядок выполнения курсового проекта

1. Согласно варианту задания разработать программу для моделирования выборки статистик критерия.

2. Смоделировать выборки статистик для разных значений n.

3. Вычислить значение $D_{n,N}$ для каждого значения n.

4. Аппроксимировать зависимость расстояния до предельного закона распределения функцией an^{-b}.

5. Определить объем выборки n^*, начиная с которого расстояние до предельного закона не превышает 0.01.

6. Проверить полученный результат, проведя контрольный эксперимент.

7. При необходимости, повторить исследования в пп. 1–6 для разных законов распределения случайной величины, способах и числе интервалов группирования, доли цензурирования.

2. Краткие теоретические сведения

2.1 Статистика критерия Жарка-Бера

Критерий (нормальности) проверки нормальности на симметричность и эксцесс [1] в англоязычной литературе называется критерием «Jarque-Bera» (Жарка-Бера) и имеет статистику [3]:

$$JB = \frac{n}{6}\left(\left(\sqrt{b_1}\right)^2 + \frac{(b_2-3)^2}{4}\right),$$

где

$$b_1 = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{\frac{1}{n}\left(\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^{3/2}}, \qquad b_2 = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\frac{1}{n}\left(\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^{2}}.$$

Критическая область критерия правосторонняя, и это можно объяснить тем, что его статистика по сути является суммой квадратов невязок между значениями выборочных оценкок параметров b_1, b_2 (коэффициентов ассиметрии и эксцесса) и теоретическими значениями параметров формы $b_1^* = 0$, $b_2^* = 3$, соответствующими нормальному закону распределения.

Асимптотически статистика подчиняется закону распределения хи-квадрат с двумя степенями свободы χ_2^2 [1]. Конкретных данных о скорости сходимости к предельному закону найти не удалось. Известно только то, что распределение статистики сходится к предельному закону очень медленно [1].

3. Результаты исследований

3.1 Выбор параметров моделирования

Параметры моделирования:

Погрешность моделирования $\delta = 0.001$.

Доверительная вероятность $P\{D < \delta\} = 0.99$.

Тогда $N = 2649147$

3.2 Моделирование распределения статистики

Так как известно, что распределение статистики сходится к предельному очень медленно [1], считается целесообразным выбрать следующие значения $n = 100, 200, 400, 800, 1600, 3200, 6400$.

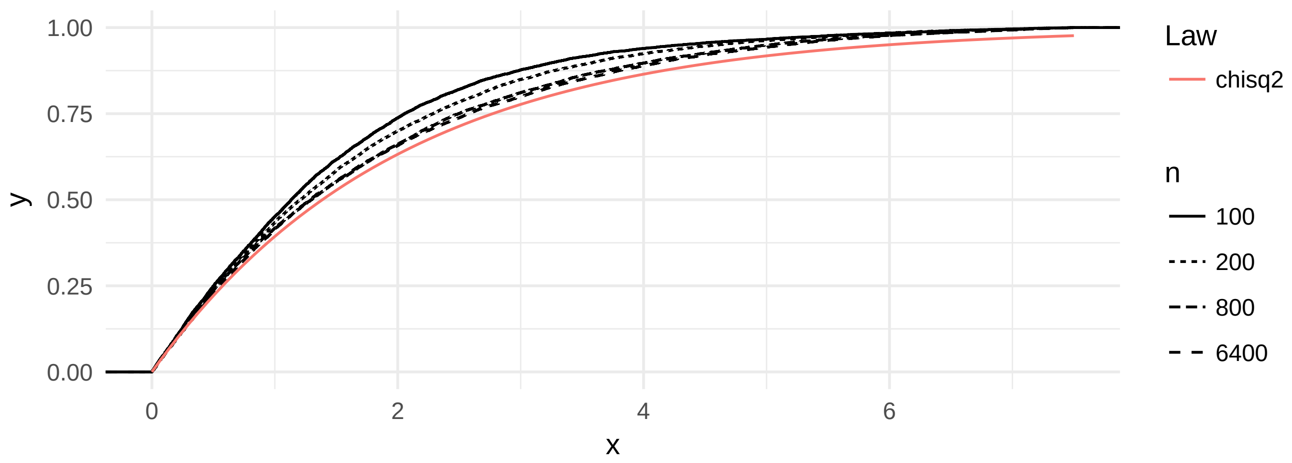

Рис. 1. Эмпирические функции распределения статистики критерия при различных объемах выборки n и предельный закон χ^2_2.

n	D_n
100	0.090865
200	0.054222
400	0.030550
800	0.017850
1600	0.009664
3200	0.004996
6400	0.002503

Таблица 1. Расстояние Колмогорова от эмпирической функции распределения до предельного закона в зависимости от объема выборки n.

3.3 Аппроксимация расстояния до предельного закона степенной функцией

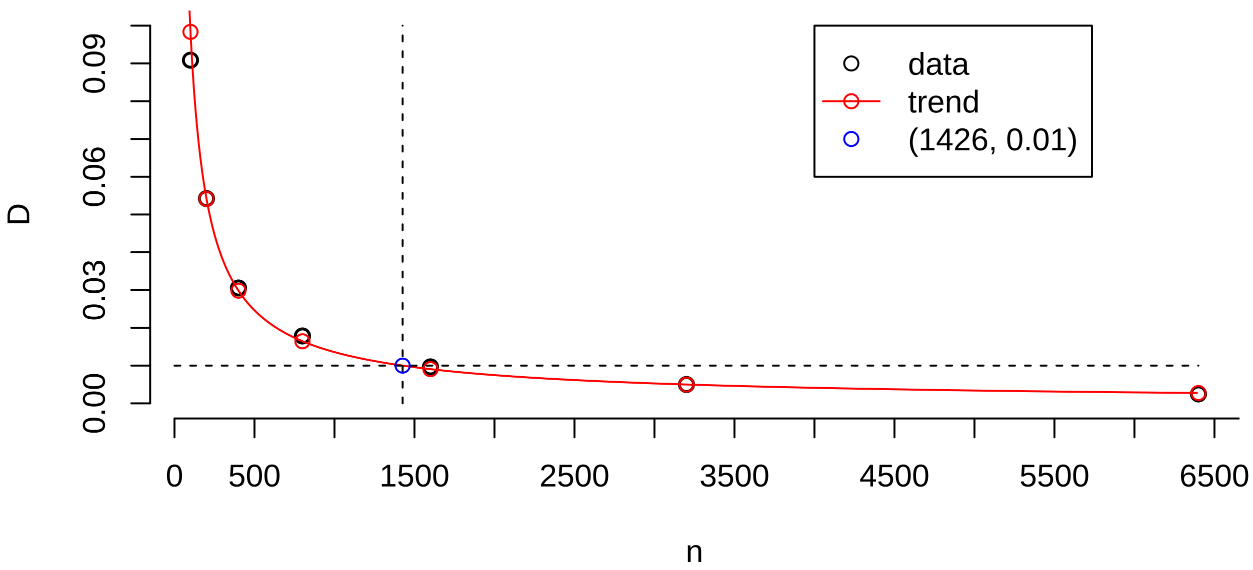

Рис. 2. Данные зависимости расстояния Колмогорова D от объема выборки n и построенный по этим данным степенной тренд.

Полученная модель:

$$D(n) = 1.642493\ n^{-0.86026}$$

Коэффициент детерминации $R^2 = 0.997339$.

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

Необходимо решить уравнение $\varepsilon = an^{-b}$ относительно n. Для этого прологарифмируем обе части уравнения. Воспользуемся десятичным логарифмом. Выразим n из полученного уравнения, в итоге получим:

$$n(\varepsilon) = \exp\left\{\frac{\ln(\varepsilon) - \ln a}{b}\right\}.$$

$$n(0.01) = 1426$$

Для проверки предсказания по модели проведем контрольный эксперимент.

3.5 Контрольльный эксперимент

В результате проведения контрольного эксперимента при $n = 1426$ расстояние от эмпирической функции распределения до предельного закона составило

$$D_{1426} = 0.010845,$$

это значение находится в пределах заданной статистической погрешности моделирования $\delta = 0.001$, т.е. $0.01 - \delta < 0.010845 < 0.01 + \delta$.

4. Заключение

В результате проведения экспериментальных исследований были смоделированы распределения статистики Жарка-Бера, результаты показали сходимость распределения к предельному, была построена экспериментальная зависимость расстояния от эмпирического распределения до предельного закона в классе степенных функций.

Результаты экспериментальных исследований хорошо согласуются с теоретическими положениями, в частности с теоремой Колмогорова, а также с другими результатами, полученными ранее другими исследователями.

Список литературы

[1] Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход : монография / Б.Ю. Лемешко, С.Б. Лемешко, С.Н. Постовалов, Е.В Чимитова. — Новосибирск : Изд-во НГТУ, 2011. — 888 с. (серия «Монографии НГТУ»).

[2] Jarque, C. M. and Bera, A. K. (1987): A test for normality of observations and regression residuals. — International Statistical Review, vol. 55, pp. 163–172.

[3] Ilya Gavrilov and Ruslan Pusev (2014). normtest: Tests for Normality. R package version 1.1. https://CRAN.R-project.org/package=normtest

[4] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[5] Pierre Lafaye de Micheaux, Viet Anh Tran (2016). PoweR: A Reproducible Research Tool to Ease Monte Carlo Power Simulation Studies for Goodness-of-fit Tests in R. Journal of Statistical Software, 69(3), 1-42. doi:10.18637/jss.v069.i03

[6] Yihui Xie (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.15.1.

Приложение А Исходные тексты программ

А.1 Фрагменты исходных текстов

```
statmod <- function(stat, n, N, h0) {
    sink(stderr())
    str <- sprintf("statmod(stat='%s', n=%s, N=%s, h0='%s')",
        stat, n, N, h0)
    message(str)

    h0 <- paste("r", h0, sep = "")

    n_grid <- rep(n, N)

    foo <- function(n) {
        # NOTE:
        x <- do.call(h0, args = list(n = n))
        do.call(stat, args = list(X = x))
    }

    library(parallel)
    cores <- detectCores() - 1
    n_grid <- split(n_grid, 1:length(n_grid)%%cores)
    jobs <- list()
    for (i in 2:cores - 1) {
        job <- mcparallel(aaply(as.matrix(n_grid[[i]]),
            c(1), foo))
        jobs <- c(jobs, list(job))
    }
    i <- length(n_grid)
    job <- mcparallel(aaply(as.matrix(n_grid[[i]]), c(1),
        foo, .progress = "text"))
    jobs <- c(jobs, list(job))
    rez <- mccollect(jobs)
    X <- unlist(rez, use.names = F)

    h0 <- substr(h0, 2, nchar(h0))

    attr(X, "n") <- n
    attr(X, "N") <- N
    attr(X, "stat") <- stat
    attr(X, "h0") <- h0

    # TODO: return numeric with attributes
    df <- data.frame(x = X, n = rep(as.factor(n), N), N = rep(as.factor(N),
```

```
        N), h0 = rep(h0, N), stat = rep(stat, N))
    sink()
    return(df)
}
```