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

Курсовой проект

на тему

«Проверка гипотезы нормальности по критерию Эппса-Палли»

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, 2016 г.

Содержание

Введение

Целью работы является ознакомление с современными тенденциями развития аппарата прикладной математической статистики и состоянием программного обеспечения задач статистического анализа. Освоение методов статистического моделирования как средства исследования и развития аппарата прикладной математической статистики. Исследование особенностей методов проверки статистических гипотез. Закрепление навыков проведения самостоятельных исследований.

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 критерий Эппса-Палли проверки принадлежности выборки нормальному закону распределения. В связи с этим и сказанным в абзаце выше предполагается считать тему работы актуальной.

1. Постановка задачи

Разработать программный модуль проверки гипотезы нормальности по критерию Эппса-Палли. Для этого следует разработать отдельно процедуру вычисления статистики критерия и процедуру вычисления достигаемого уровня значимости при проверки гипотезы методом статистических испытаний Монте-Карло.

1.1 Статистика критерия

Критерий Эппса-Палли основан на сравнении эмпирической (выборочной) и теоретической характеристических функций, выражение статистики данного критерия имеет вид

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

где $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$, $\hat{\mu}_2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$.

Точнее, это есть взвешенный интеграл модуля разности теоретической (нормального распределения) и эмпирической (выборочной) характеристических функций.

Гипотезу о нормальности отвергают при больших значениях статистики. Точный предельный закон распределения данной статистики неизвестен, но он может быть хорошо аппроксимирован бета-распределением 3-его рода [1].

1.2 Достигаемый уровень значимости

Алгоритм 1. Вычисление достигаемого уровня значимости статистического критерия методом Монте-Карло.

Входные данные: гипотеза H_0, выборка X_n, количество повторений N, функция вычисления статистики $S(X_n)$.

Действия:

1. Вычислить $S = S(X_n)$ – статистику критерия по выборке.
2. Установить $m = 0$.
3. Сгенерировать выборку Y_n при верной гипотезе H_0.
4. Вычислить значение $S(Y_n)$.
5. Если критическая область правосторонняя и $S(Y_n) > S(X_n)$, то $m = m+1$.
6. Повторять шаги 3-5 N раз.

Выходные данные: оценка достигаемого уровня значимости (p-value) равна

- $\hat{p} = \frac{m}{N}$ для правосторонней критической области;
- $\hat{p} = 1 - \frac{m}{N}$ для левосторонней критической области;
- $\hat{p} = 2\min(\frac{m}{N}, 1 - \frac{m}{N})$ для двусторонней критической области.

2. Аналитический обзор

Пусть $X_1, \ldots, X_n$ — случайная выборка из закона распределения $F(x)$, тогда эмпирическая характеристическая функция определяется как $\phi_n(t) = n^{-1} \sum_j \exp(itX_j)$, где t произвольный вещественный параметр. Можно показать, что $\phi_n(t)$ почти наверное сходится к теоретической характеристической функции $\phi(t)$ (характеристической функции генеральной совокупности), также между $\phi(t)$ и $F(x)$ существует взаимооднозначное соответствие, всё дает возможность использовать $\phi_n(t)$ для статистических выводов. Так критерий Эппса-Палли использует $\phi_n(t)$ для проверки сложной гипотезы о том, что $F(x)$ — нормальный закон распределения.

Пусть $\phi_0(th) = \exp(it\mu - \frac{1}{2}t^2\sigma^2)$ — характеристическая функция при верной нулевой (основной) гипотезе, μ и σ^2 — неизвестные параметры сдвига и масштаба. Статистика рассматриваемого в данной работе критерия основана на взвешенном интеграле квадрата модуля разности $\phi_n(t) - \hat{\phi}_0(t)$, где $\hat{\phi}_0(t)$ зависит от μ и σ^2, оцененных по выборке.

Хифкот (Heathcote, 1972) и Фейджин и Хифкот (Feigin and Heathcote, 1977) ранее рассмотрели использование отдельно вещественной и мнимой составляющих $\phi_n(t)$ для проверки простой гипотезы. Их критерий основан на том факте, что, для заданного t, обе составляющие $\phi_n(t)$ а асимптотике подчиняются нормальному закону распределения, с некоторым параметром сдвига. Если альтернативная гипотеза также является простой, то можно найти значение t, такое, что мощность критерия будет максимальна для выборок с большим объемом. Эппс, Синглтон и Палли (Epps, Singleton & Pulley, 1982) показали как можно использовать выборочную производящую функцию моментов для проверки сложной гипотезы о том, что данные подчиняются одному из двух заданных семейств распределений.

Мурота и Такеучи (Murota & Takeuchi, 1981) недавно (на момент публикации [2]), критерий инвариантный относительно параметров сдвига и масштаба, основанный на статистике $\tilde{a}_n(t) = |\phi_n(t/S)|^2$, где S — выборочное стандартное отклонение. Применительно к задаче проверки на нормальность, критерий на основе $\tilde{a}_n(t)$ показал высокую мощность при $t \approx 1.0$ относительно шести семейств симметричных распределений, выступавших в качестве альтернативных гипотез. Но для относительно альтернатив, при которых распределения являлись ассиметричными, критерий не показал высокую мощность.

В попытках разработать универсальный критерий нормальности в работе [2] был тщательно исследован критерий со статистикой

$$M(t) = \frac{1}{n}\sum_j \exp\{t(X_j - \bar{X})/S\},$$

которая является производящей функцией моментов нормированной выборки.

Авторы выяснили, что, особенно для ассиметричных распределений, мощность критерия значительно меняется в зависимости от t, и ни одно фиксированное значение t не может дать удовлетворительный универсальный критерий.

Было предложено несколько критериев, основанных на разных значениях t. Используя свойство совместной нормальности вещественной и мнимой частей $\phi_n(.)$, вычисленных при каждом из выбранных значений t, Кутровелис (Koutrouvelis, 1980) и Кутровелис и Келлермейер (Koutrouvelis & Kellermeier, 1981) построили статистику, распределенную в асимптотике как хи-квадрат при верной нулевой (основной) гипотезе. Хифкот (Heathcote, 1972) предложил критерий, основанный на супремуме по t от $|\phi_n(t) - \phi_0(t)|$. Хотя t является непрерывным параметром, супремум будет определяться относительно некоторого конечного множества, $(t_1, \ldots, t_J)$. Эта процедура интуитивно логична, но с её применением связаны некоторые трудности, подобные трудностям применения критерия типа хи-квадрат. Дело в том, что точно так же как и не существует одного верного способа выбора граничных точек разбиения на интервалы для критерия типа хи-квадрат, так же и здесь не существует способа выбора множества значений параметра t, такого чтобы критерий оставался наиболее мощным для как можно большего числа альтернатив. Выбор этого множества путем проб и ошибок также можно считать безнадежным занятием. Более рациональный подход предложили Ферверджер и Мureика (Feuerverger & Mureika, 1977), эти авторы предложили использовать статистику вида $\int\{Im\ \phi_n(t)\}^2 dG(t)$, где пределы интегрирования $(-\infty, +\infty)$, для проверки симметичности $F(x)$. Здесь весовая функция $G(t)$ сама является по сути функцией распределения с производной $G'(t)$ симметричной относительно нуля. В зависимости от выбора $G(t)$, эта статистика может повторять поведение $\phi_n(t)$ для любых значений t [2].

Статистика предложенного критерия проверки сложной гипотезы о нормальности основана на интеграле

$$\int_{-\infty}^{+\infty} \left|\phi_n(t) - \hat{\phi}_0(t)\right|^2 dG(t), \tag{1}$$

где $\hat{\phi}_0 = \exp(it\bar{X} - \frac{1}{2}t^2S^2)$, $\bar{X} = n^{-1}\sum X_j$ и $S^2 = n^{-1}\sum(X_j - \bar{X})^2$.

При выборе весовой функции $G(t)$ следует руководствоваться следующими соображениями. Во-первых, она должна увеличиваться при тех значениях t, при которых значение $|\phi_1(t) - \phi_0(t)|^2$ велико, $\phi_1(t)$ относится к альтернативной гипотезе. Многие непрерывные распределения могут придавать большие значения $|\phi_1(t) - \phi_0(t)|^2$ на интервале $0 < t < 3$, если эти распределения записать в стандартной форме. Во-вторых, $G(t)$ должна принимать большие значения там, где статистика $\phi_n(t)$ является относительно точной оценкой $\phi(t)$. Можно показать, что

$$E\{|\phi_n(t) - \phi(t)|^2\} = n^{-1}\{1 - |\phi(t)|^2\}.$$

Так как $|\phi(0) = 1|$ и для любого непрерывного распределения $\lim|\phi(t)| = 0$ при $t \to 0$, то видно, что точность эмпирической характеристической функции как оценки теоретической наибольшая вблизи нуля и сильно уменьшается при удалении от нуля. Тогда получается, что $G(t)$ должна увеличиваться на интервале вблизи нуля, включая нуль. Из определений $\phi_n(t)$ и $\phi(t)$ понятно, что длина такого интервала должна быть обратнопропорционально зависеть от дисперсии выборки. И последнее практическое соображение по выбору $G(t)$ заключается в том, что $G(t)$ должна быть такой, чтобы интеграл (1) был существовал.

Обозначим $dG(t) = g(t)dt$ и рассмотрим функцию

$$g(t) = \left\{\alpha S/\sqrt{(2\pi)}\right\}\exp\left(-\frac{1}{2}\alpha^2S^2t^2\right), \tag{2}$$

которая является нормальным распределением с нулевым параметром сдвига и параметром масштаба $(\alpha S)^{-2}$. Итак, первые два тревования к весовой функции выполнены, при условии правильного выбора α; более того, (1) существует:

$$T(\alpha) = n^{-2}\sum_{j=1}^{n}\sum_{k=1}^{n}\exp\left\{-\frac{1}{2}(X_j - X_k)^2/(\alpha^2S^2)\right\}$$
$$- 2n^{-1}(1+\alpha^{-2})^{-\frac{1}{2}}\sum_{j=1}^{n}\exp\left[-\frac{1}{2}(X_j - \bar{X})^2/\left\{S^2(1+\alpha^2)\right\}\right] + (1+2\alpha^{-2})^{-\frac{1}{2}}. \tag{3}$$

В [1] по умолчанию $\alpha = 1$. Значение этого параметра несколько влияет на мощность критерия относительно различных альтернатив. Это может быть

полезно при наличии априорных знаний о некоторых свойствах наблюдаемого закона [2].

3. Результаты исследований

Параметры моделирования:

— Число повторений Монте-Карло $N = 16600$.

3.1 Моделирование распределения статистики

Было проведено моделирование распределения статистики при верной основной гипотезе (стандартный нормальный закон) при разных объемах выборок n в системе статистического анализа одномерных наблюдений ISW [3] и в среде программирования и статистического анализа R [4] с использованием пакета PoweR [5]. Построены сравнительные графики эмпирических функций распределений статистики.

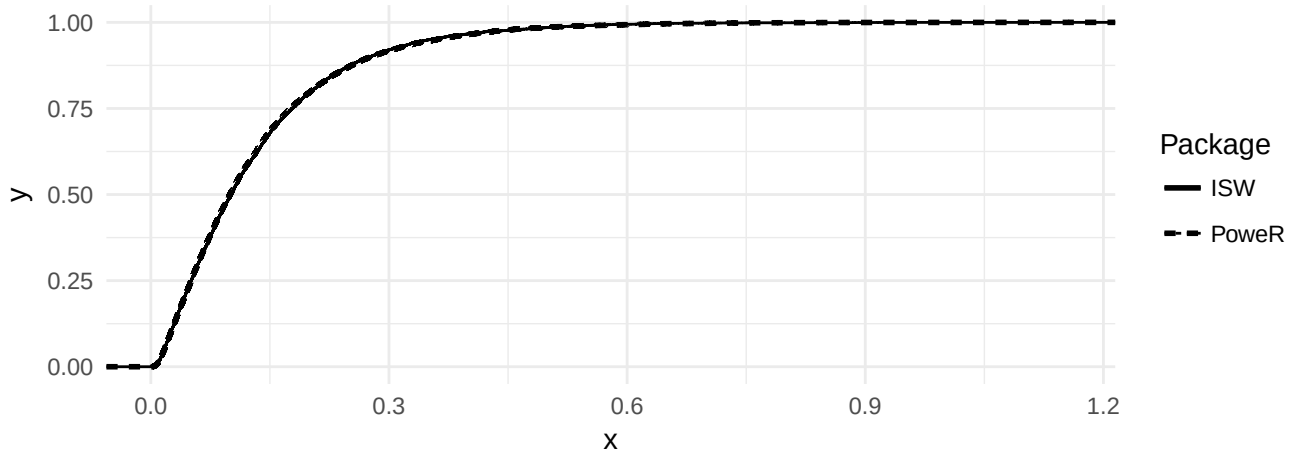

Рис. 1. Эмпирическая функция распределения статистики Эппса-Палли, объем выборки $n = 10$, число повторений Монте-Карло $N = 16600$.

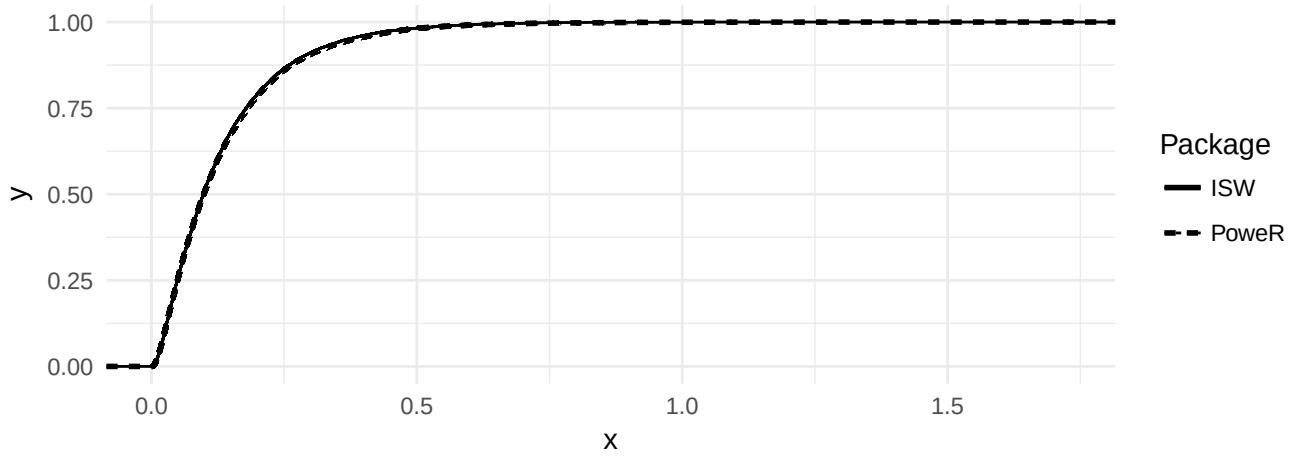

Рис. 2. Эмпирическая функция распределения статистики Эппса-Палли, объем выборки $n = 50$, число повторений Монте-Карло $N = 16600$.

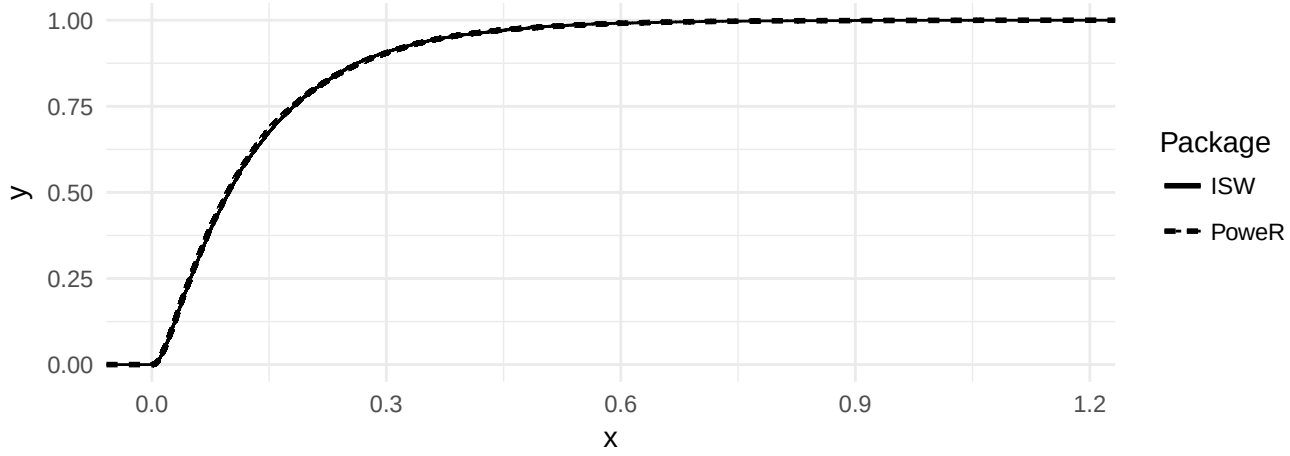

Рис. 3. Эмпирическая функция распределения статистики Эппса-Палли, объем выборки $n = 200$, число повторений Монте-Карло $N = 16600$.

3.2 Проверка гипотезы о нормальности

Были проведены процедуры проверки гипотез о нормальности выборок, сгенерированных по различным законам распределения, т.е. при справедливости различных альтернативных гипотезах. Объем выборки $n = 200$, число испытаний Монте-Карло при вычислении p-value $N = 16600$.

Таблица 1. Результаты проверки нормальности с помощью PoweR.

Закон распределения	Значение статистики	p-value
Нормальный(0,1)	0.1279377	0.3850000
Лапласа(0,1)	1.3119339	0.0000000
Логистический(0,1)	0.5363731	0.0140361
Обобщ. нормальный(4,0,1)	0.4626721	0.0271687

Таблица 2. Результаты проверки нормальности с помощью ISW.

Закон распределения	Значение статистики	p-value
Нормальный(0,1)	0.1279390	0.1000000
Лапласа(0,1)	1.3119400	0.0000000
Логистический(0,1)	0.5363740	0.0100000
Обобщ. нормальный(4,0,1)	0.4626730	0.0250000

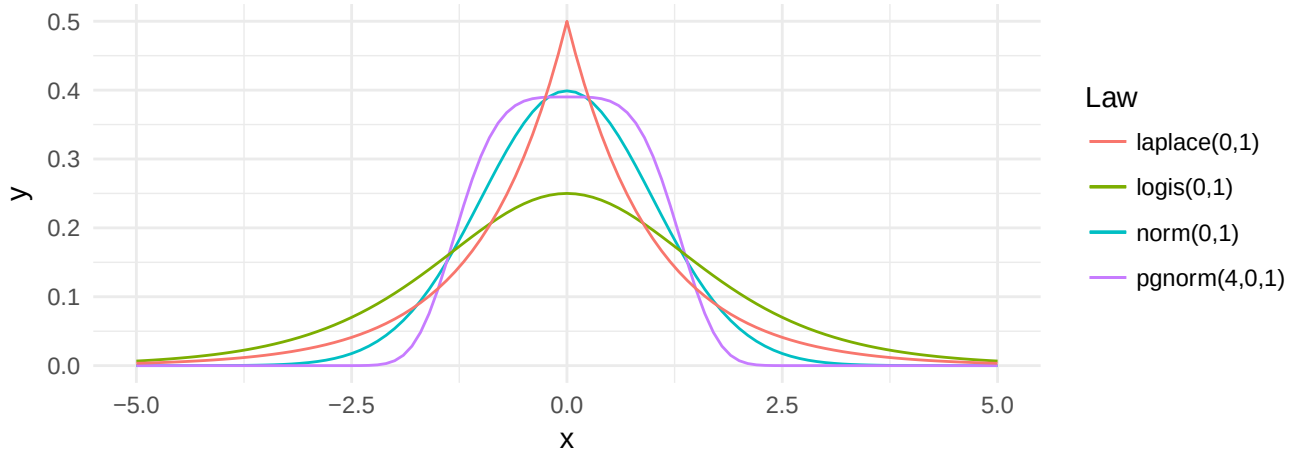

Рис. 4. Теоретические плотности рассматриваемых законов распределений.

3.3 Проверка нормальности реальных наблюдений

3.3.1 Описание набора данных

Была проведена процедура проверки нормальности реальных наблюдений из набора данных ”US Arrests” — число арестов в различных штатах США за

различные преступления (разбойные нападения, изнасилования, убийства) за 1973 год. Также имеется массив относительного количества городского населения в этих штатах.

3.3.2 Постановка задачи

Проверим относительное количество городского населения в различных штатах США (1973 год) на нормальность.

3.3.3 Результаты PoweR

Приведены результаты при различном числе испытаний Монте-Карло N.

Значение статистики $T_{EP} = 0.154636$

Количество повторений Монте-Карло $N = 100$

- Значение p-value: 0.26

Количество повторений Монте-Карло $N = 8400$

- Значение p-value: 0.308214

Количество повторений Монте-Карло $N = 16600$

- Значение p-value: 0.311747

Количество повторений Монте-Карло $N = 33200$

- Значение p-value: 0.313163

3.3.4 Результаты проверки гипотезы в ISW

Значение статистики: 0.154636

p-value: 0.1

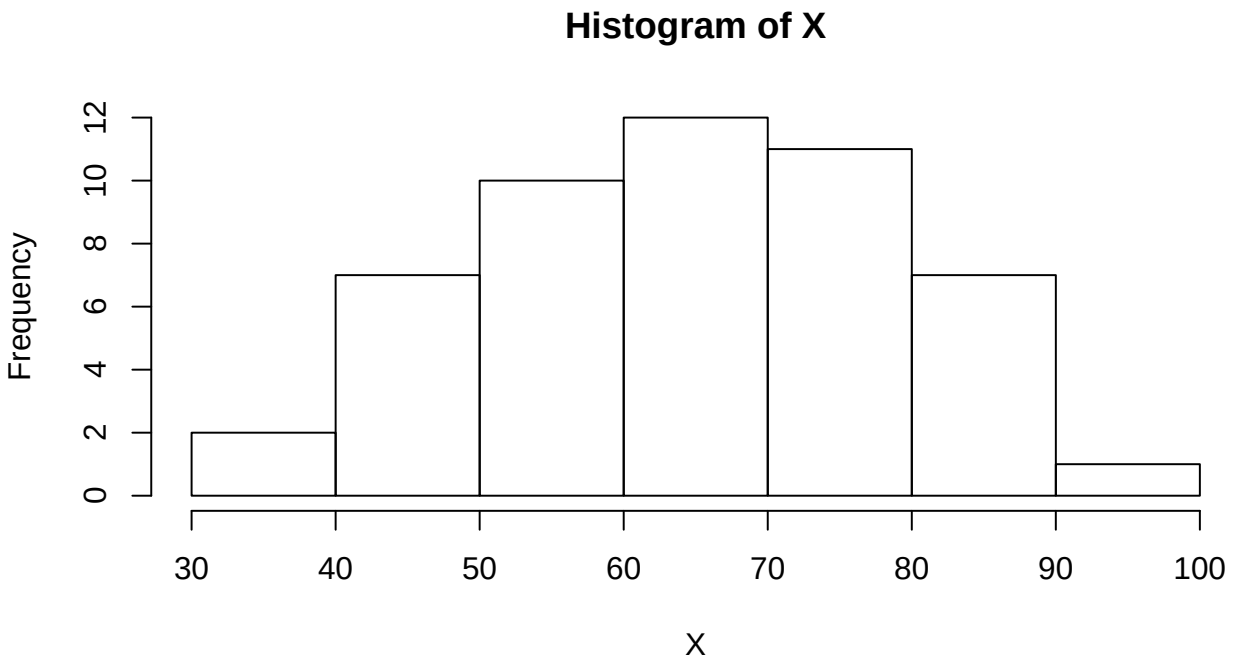

Рис. 5. Частотная гистограмма относительного количества городского населения в различных штатах США.

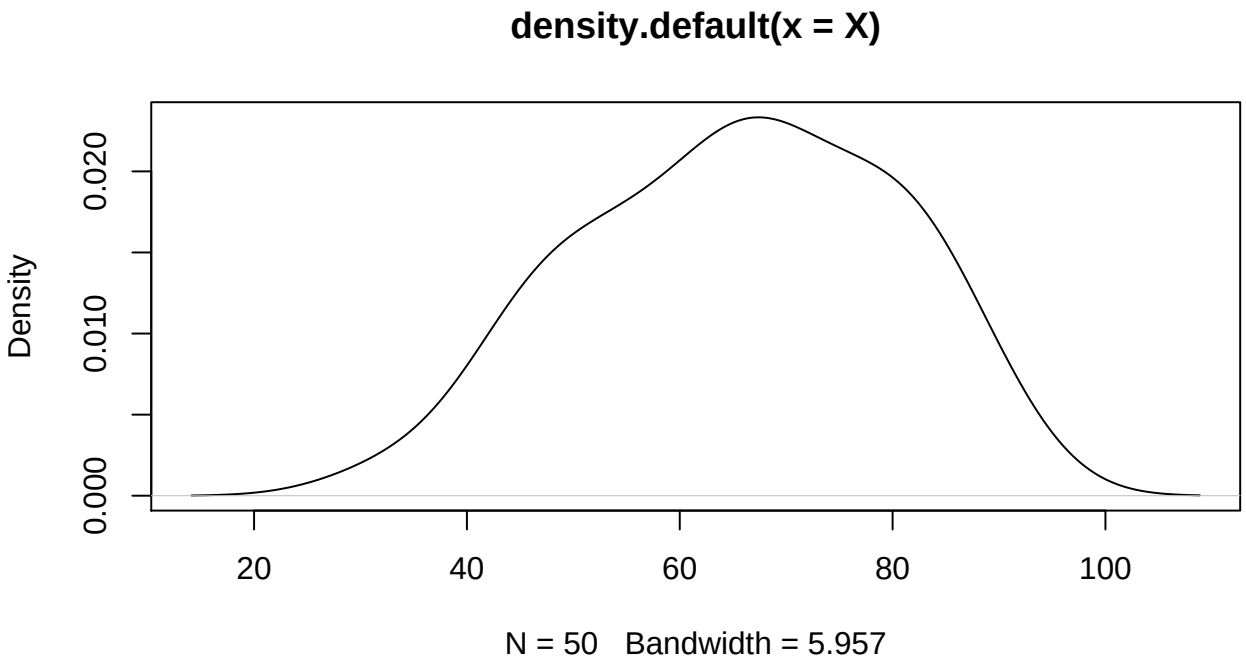

Рис. 6. Ядерная оценка плотности распределения относительного количества городского населения в различных штатах США.

4. Заключение

В подразделе 3.1 представлены результаты моделирования эмпирического распределения статистики с помощью R (PoweR) и ISW. Результаты хорошо согласуются. В связи с этим предполагается считать, что соответствующие алгоритмы реализованы корректно.

По результатам проверки нормальности с помощью ISW видно, что p-value вычисляется по таблице квантилей эмпирического распределения статистики.

В подразделе 3.2 проведено моделирование достигаемого уровня значимости при проверки гипотезы о нормальности при справливости различных альтернатив и основной гипотезы.

В общем наблюдается следующая тенденция: чем больше значение статистики, тем меньше значение уровня значимости, что логично, т.к. статистика критерия является мерой расстояния между эмпирической и характеристической функциями распределений, а значит и между законами распределений — тоже.

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] Epps T. W. A test for normality based on the empirical characteristic function / T. W. Epps, L. B. Pulley // Biometrika. – 1983. – Vol. 70. – P. 723–726.

[3] Лемешко Б.Ю., Постовалов С.Н. Система статистического анализа наблюдений и исследования статистических закономерностей // Материалы международной НТК "Информатика и проблемы телекоммуникаций". — Новосибирск, 2001. — С. 80 – 81.

[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

```
Tep <- function(X) {
    exp(-(.C(dontCheck("stat31"), as.double(X), as.integer(length(X)),
        as.double(c(0.05, 0.1)), as.integer(2), rep(" ",
            50), 0L, statistic = 0, pvalcomp = 1L, pvalue = 0,
        cL = as.double(0), cR = as.double(0), as.integer(0),
        alter = as.integer(0), decision = as.integer(rep(0,
            2)), stat.pars = as.double(1), as.integer(1),
        PACKAGE = "PoweR")$statistic))
}
mypvalMC <- function(X, N, law.index = 2, law.pars = NULL,
    Rlaw = NULL) {
    S <- TepMC(length(X), law.index, 31, M = N, law.pars = law.pars,
        Rlaw = Rlaw)
    m <- length(S[which(S > Tep(X))])
    m/N
}
TepMC <- function(...) {
    exp(-(compquant(...)$stat))
}
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) {
```

```
        x <- do.call(h0, args = list(n = n))
        do.call(stat, args = list(X = x))
    }

    X <- aaply(as.matrix(n_grid), c(1), foo, .progress = "text")

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

    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)
}
n <- 10
dfp <- statmod("Tep", n, N, "norm")

x <- read.table(file.path("data", "eppspulley10.dat"), skip = 2,
    col.names = "x")
dfi <- data.frame(x = x, n = n)
plot.ecdfs <- function(dfp, dfi) {
    cap <- sprintf("Эмпирическая функция распределения статистики
                    Эппса-Палли, объем выборки $n = %s$, число повторений
                    Монте-Карло $N = %s$.",
        dfp$n[1], dfp$N[1])

    assign("caption", cap, envir = .GlobalEnv)

    ggplot(dfi, aes(x = x, linetype = "ISW")) + stat_ecdf(size = 0.5) +
        stat_ecdf(data = dfp, mapping = aes(x = x, linetype = "PoweR"),
            size = 1) + scale_linetype(name = "Package")
}
plot.ecdfs(dfp, dfi)
X <- gensample(2, n)$sample
S <- Tep(X)
pval <- mypvalMC(X, N, 2)
law <- sprintf("Нормальный(0,1)")

dfp <- data.frame(law = law, S = S, pval = pval)
X <- gensample(1, n)$sample
S <- Tep(X)
pval <- mypvalMC(X, N, 2)
law <- sprintf("Лапласа(0,1)")

dfa <- data.frame(law = law, S = S, pval = pval)
dfp <- rbind(dfp, dfa)
```