Mathematics in Modern Technical University

Mathematics in Modern Technical University, 2018(1), 91–100
http://mmtu.matan.kpi.ua

Методи знаходження законів розподілів випадкових величин за даними статистичних вибірок засобами мови R

О. О. Диховичний, Н. В. Круглова, О. І. Вирстюк

Кафедра математичного аналізу та теорії ймовірностей,
КПІ ім. Ігоря Сікорського, Київ, Україна

a.dyx@ukr.net

Анотація

У статті досліджено методи підбору теоретичного ймовірнісного розподілу для змодельованих статистичних даних засобами мови статистичного програмування R. Розглянуто графічні засоби підбору закону розподілу: побудова гістограм, емпіричних і теоретичних щільностей і функцій розподілу, P-P і Q-Q діаграм. Досліджено функції оцінювання параметрів законів розподілу методами: моментів, квантілів, найбільшої вірогідності та найменшої відстані. Перевірено гіпотези про закон розподілу за допомогою критерію Колмогорова–Смірнова, а також критеріїв AIC, BIC.

Відповідний підбір, як приклад застосування, проведено для зімітованого за спеціальним авторським алгоритмом розподілу максимуму звуження поля Ченцова на певну криву засобами мови R.

Ключові слова: мова R; поле Ченцова; ґаусівський процес; метод моментів; метод квантілів; метод найбільшої вірогідності; метод найменшої відстані; критерії згоди.

MSC2010 62E04

УДК 519.233.33

Cite as: Dykhovychnyi, O. O., Kruglova, N. V., Virstiuk, O. I. (2018). Methods for identification of probability distribution of random variables from data samples with R statistical computing language. *Mathematics in Modern Technical University*, 2018(1), 91–100.

1 Вступ і попередні відомості

У ряді статистичних досліджень знаходження точного розподілу певного функціоналу від випадкового процесу є достатньо складною а, іноді, й нерозв'язною задачею. Прикладом такої задачі є знаходження розподілів функціоналів від поля Ченцова. Але в деяких випадках, на підставі відомих імовірнісних характеристик процесу (середнє, коваріаційна функція, моменти тощо), стає можливим змоделювати вибірку, яка відповідає генеральній сукупності із заданими характеристиками. А далі, емпіричним шляхом підібрати найбільш підходящий ймовірнісний розподіл. Таке дослідження можна побудувати за наступною схемою:

1) підбір можливого теоретичного закону розподілу, який найкраще описує вибірку;

2) обчислення оцінок основних параметрів розподілу;

3) перевірка гіпотези про узгодженість емпіричного та теоретичного розподілів.

Якщо припустити, що випадкова величина є абсолютно неперервною, то для знаходження закону розподілу достатньо підібрати вигляд щільності розподілу. Найбільш поширеним методом знаходження щільності розподілу за даними статистичної вибірки є *метод гістограм* (Syzrantsev, Nevelev, & Golofast, 2006), (Wolverton & Wagner, 1969). Недоліками цього методу є низька надійність і нестійкість до викидів. Також можна застосовувати такі методи: Парзена–Розенблата (Lapko, Chentsov, Krokhov, & Feldman, 1996), інтегральної оцінки щільності, стохастичної регулярізації та інші.

Після припущення про вигляд щільності розподілу, визначають параметри щільності розподілу (якщо це можливо). Подаймо перелік найвідоміших методів знаходження оцінок параметрів розподілу.

Метод моментів (Cramér, 1946). Нехай $x_1, x_2, ..., x_n$ — вибірка з розподілу $F(x, \theta_1, \theta_2, ..., \theta_s)$. Потрібно одержати оцінки для невідомих параметрів $\theta_1, \theta_2, ..., \theta_s$. Суть методу полягає у прирівнюванні певної кількості вибіркових моментів $\tilde{m}_k$ відповідним теоретичним моментам

$$m_k = \int_{-\infty}^{\infty} x^k f(x, \theta_1, \theta_2, ..., \theta_S) dx.$$

Кількість рівнянь відповідає кількості параметрів, які визначають щільність розподілу.

Метод максимальної вірогідності (Cramér, 1946). Це метод оцінювання параметрів, який ґрунтується на максимізації функції вірогідності вибірки. Якщо вибірка має неперервний розподіл, то функцію вірогідності описують сумісною щільністю розподілу:

$$L(x_1; x_2; ...; x_n; \theta) = f(x_1; \theta) f(x_2; \theta) ... f(x_n; \theta).$$

Оцінками максимальної вірогідності є значення параметра θ, які максимізують функцію L. Часто простіше шукати максимум функції $\ln L$, який збігається з максимумом функції L завдяки монотонності логарифма.

Метод порядкових статистик (метод квантілів) (Kobzar, 2006). Цей метод дуже схожий на метод моментів: вибирається така ж кількість квантілів, скільки невідомих параметрів необхідно оцінити. Потім теоретичні квантілі, які виражені через параметри розподілу, прирівнюють до емпіричних квантілів. Розв'язок відповідної системи є оцінками невідомих параметрів. Ефективність оцінок, одержаних таким методом, не вища за метод моментів.

Метод найменшої відстані (Wolfowitz, 1957). Знаходять відстані між емпіричною та теоретичною функціями розподілу в різних метриках мінімізують їх. Використовують, зокрема, метрики: Крамера – фон Мізеса, Колмогорова–Смірнова, Андерсона–Дарлінґа.

2 Результати дослідження

Як зазначено вище, запропонована схема підбору найкращого розподілу, передбачає наступні етапи з застосуванням пакетів програмного середовища R (*The Comprehensive R Archive Network*, n.d.):

1) підбір теоретичної щільності розподілу здійснюється графічним методом, а саме: за допомогою побудови гістограми й теоретичної щільності, порівнянні емпіричної і теоретичної функцій розподілу, побудові P-P і Q-Q діаграм (пакет `fitdistrplus`, функції `plot`);

2) обчислення оцінок параметрів щільності (пакет `fitdistrplus`, функції `fitdist, mledist`);

3) перевірка гіпотези про узгодженість теоретичного й емпіричного розподілів (пакети `fitdistrplus` та `stats`, функції `ks.test fitdist, mledist`).

Застосуємо запропоновану схему досліджень для знаходження розподілу максимуму від ґаусівських процесів на основі статистичних вибірок, одержаних шляхом моделювання значної кількості реалізацій випадкового процесу. Швидкодія методу Холецького складає $O(n^3)$, методу розкладу за ортогональними системами — $O(n^2)$, що ускладнює моделювання достатньої кількості реалізацій процесів. Тому скористаємось алгоритмом, описаним в (Dykhovychnyi & Kruglova, 2018). У цій роботі запропоновано новий алгоритм моделювання ґаусівського процесу $Y(t)$ з нульовим математичним сподіванням і коваріаційною функцією вигляду:

$$E\left[Y(s)Y(t)\right] = u(s)v(t),\ \ s \leqslant t.$$

Цей алгоритм було застосовано для знаходження розподілу максимуму звуження двопараметричного поля Ченцова $X(s,t)$ (Chentsov, 1956), (Yeh, 1960) на криву

$$L = \left\{(s,t)|t = 1 - s^2, s \in [0,1)\right\}.$$

Точний ймовірнісний розподіл для максимуму від такого ґаусівського процесу не знайдено. Використовуючи теорему, подану в (Park & Paranjape, 1974), можна одержати інтегро-диференціальне рівняння, розв'язком якого буде шуканий розподіл. Але таке рівняння є надзвичайно складним і його точний розв'язок не знайдено. Тому доречним буде змоделювати звуження поля Ченцова на криву L і знаходити емпіричний розподіл для описаної вище випадкової величини. Алгоритм, описаний у (Dykhovychnyi & Kruglova, 2018), можна застосовувати для моделювання такого процесу, оскільки процес $X_L(s) = X\left(s, 1 - s^2\right)$ має нульове математичне сподівання й коваріаційну функцію:

$$E\left[X_L(s)X_L(t)\right] = s(1 - t^2),\ s \leqslant t.$$

Тоді функції $u(s) = s$ і $v(t) = 1 - t^2$ справджують умови застосовності цього алгоритму. Було змодельовано 10^5 реалізацій ґаусівського процесу $X_L(s)$ для одержання інформативної репрезентативної вибірки й побудовано емпіричний розподіл випадкової величини:

$$\max_{s\in[0,1)} X_L(s).$$

Для цієї вибірки використано графічний метод знаходження теоретичного закону розподілу випадкової величини, описаний вище. На рис. 1–4 побудовано емпіричні гістограми разом з теоретичними щільностями, а також Q-Q, P-P діаграми.

Спершу підбір розподілу був здійснений за допомогою функцією `fitdist`. Було перевірено всі неперервні розподіли, які передбачені функцією (нормальний, експоненціальний, гамма-розподіл, логістичний). Як видно з поданих рисунків найбільш підходящими є нормальний та гамма-розподіли.

Оцінки параметрів щільностей розподілів одержано за допомогою пакету `fitdistrplus` мови R методами: максимальної вірогідності (параметр mle), моментів (параметр mme), квантілів (параметр qme), найменшої відстані (параметр mge). У табл. 1 та табл. 2 подано оцінки параметрів для нормального й гамма-розподілів і результати перевірки гіпотез про адекватність розподілів на підставі критеріїв AIC, BIC, Коломогорова–Смірнова.

Як видно з поданих таблиць, відповідні параметри критеріїв указують на відсутність узгодженості емпіричного та теоретичного розподілів. Отже, пошук було продовжено серед інших розподілів. Для цього було використано функцію `mledist`. Ця функція дозволила перевірити на узгодженість: логнормальний розподіл, розподіл Ґумбеля (Gumbel distribution), розподіл Вейбула (Weibull distribution), оцінити їхні параметри методом максимальної вірогідності. За допомогою функції `mledist` був вибраний розподіл Вейбула. Графічний метод дав можливість не відкидати гіпотезу про такий розподіл.

За допомогою критерію Колмогорова–Смірнова було перевірено гіпотезу про узгодженість емпіричного розподілу й розподілу Вейбула з оціненими параметрами. Отже, прийнято гіпотезу про розподіл Вейбула для даної вибірки.

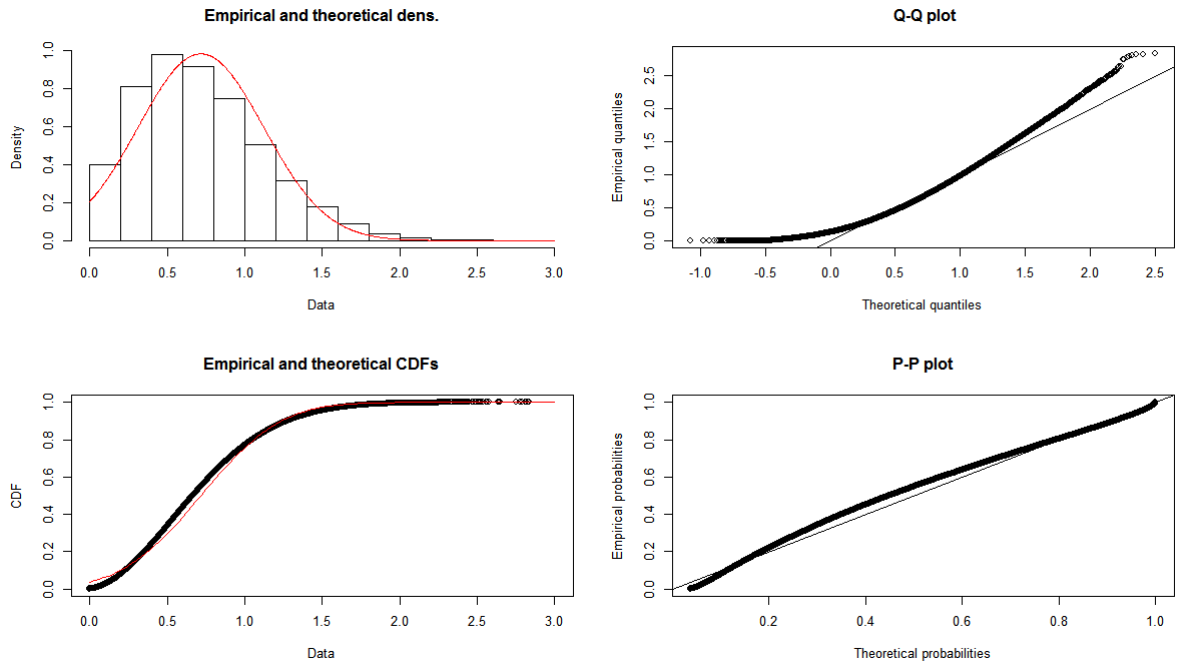

Рис. 1: Нормальний розподіл

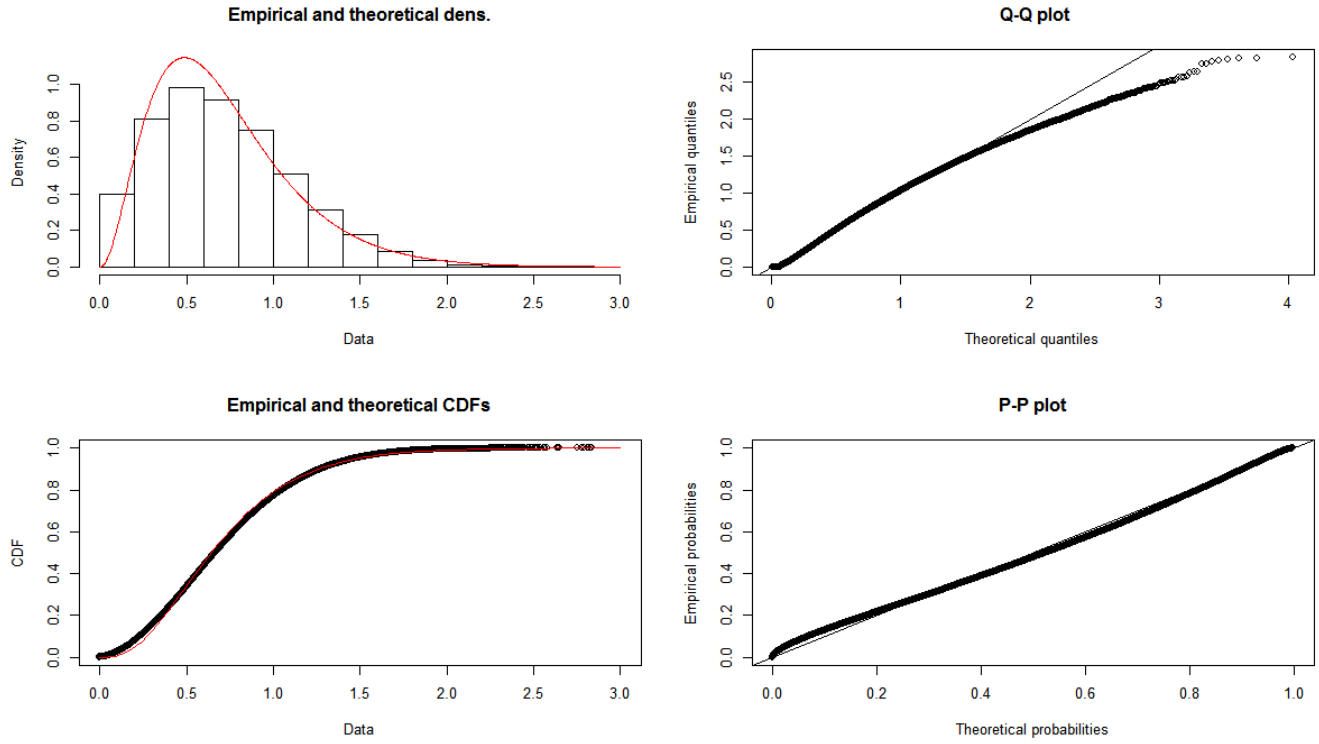

Рис. 2: Гамма-розподіл

На рис. 5 та 6 зображено гістограму, щільність теоретичного розподілу, емпіричну й теоретичну функцію розподілу, Q-Q і P-P діаграми для розподілів Гумбеля (як найкращого серед відкинутих розподілів) і Вейбула.

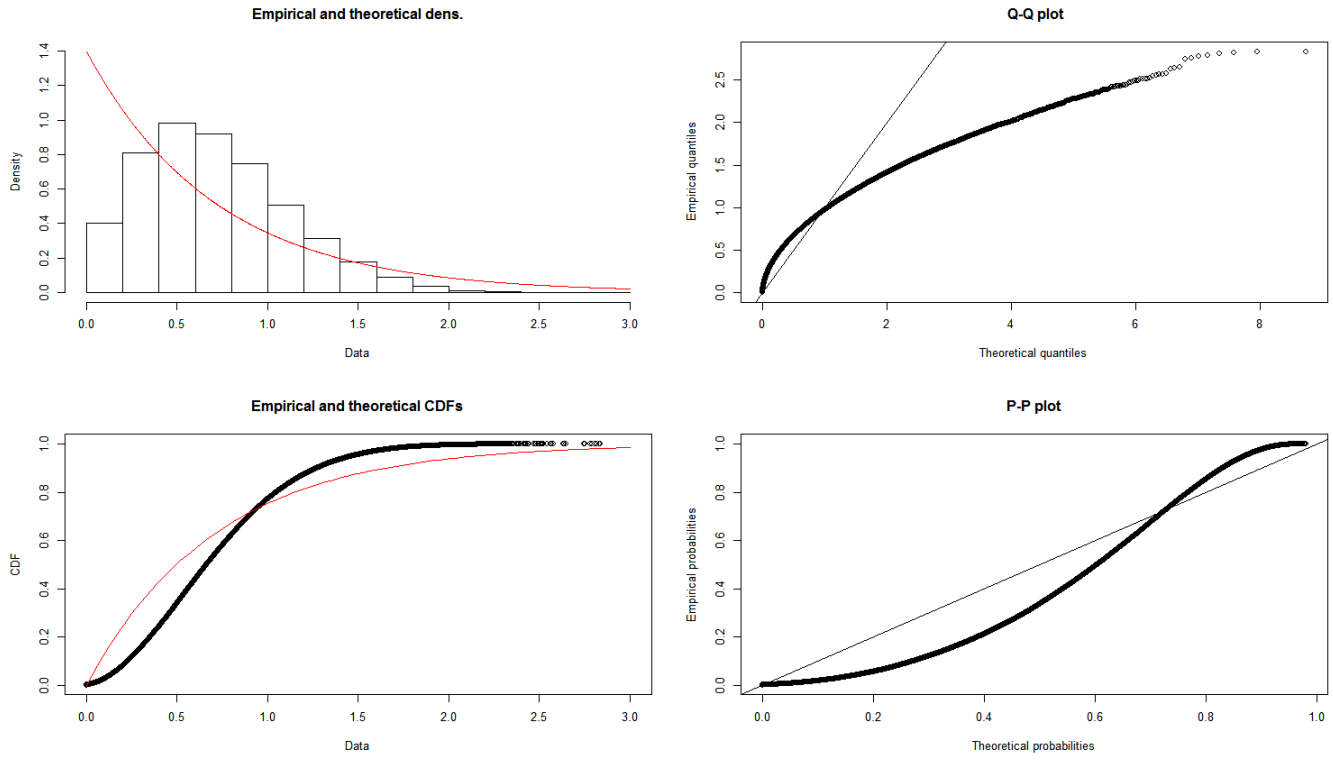

Рис. 3: Гамма-розподіл

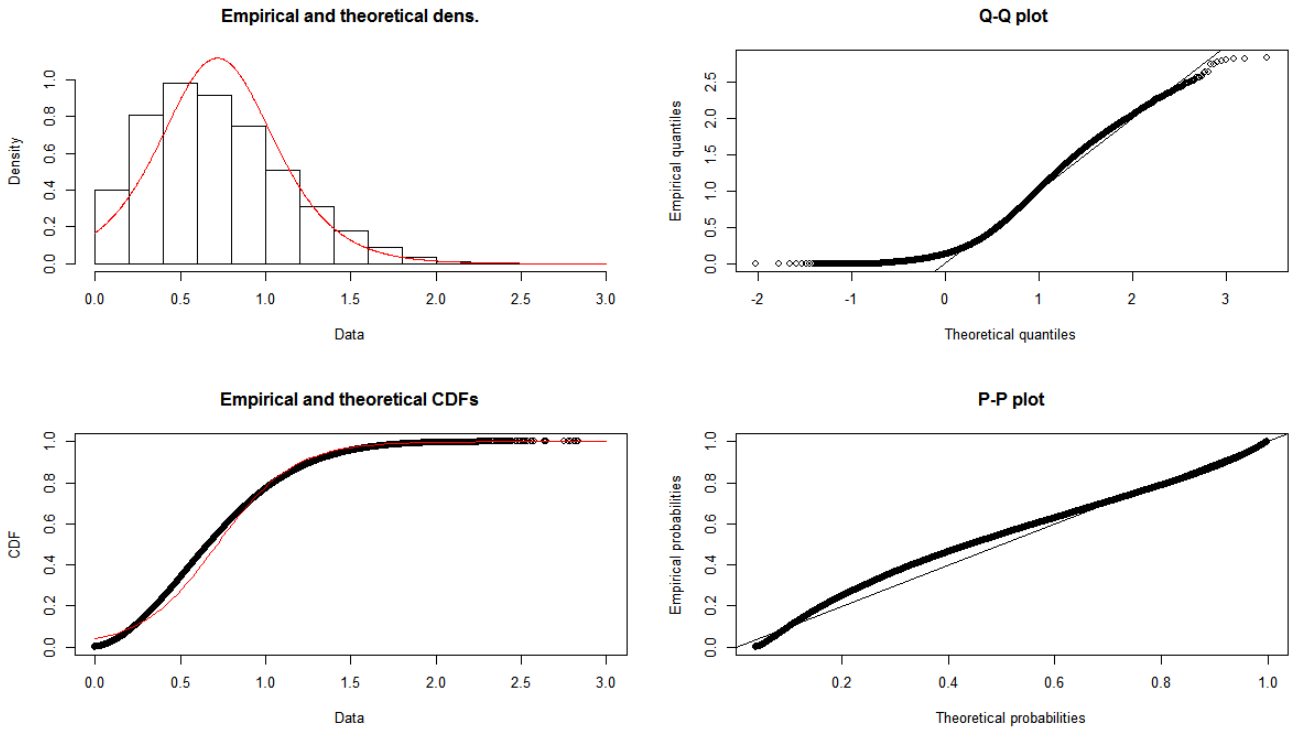

Рис. 4: Експоненціальний розподіл

У табл. 3 наведено результати оцінок відповідних параметрів та перевірки

Табл. 1: Нормальний розподіл

	Методи одержання оцінок			
	mme	mle	qme	mge
Параметри розподілу	Mean= 0.717 Sd= 0.41	Mean= 0.717 Sd= 0.41	Mean= 0.6645 Sd= 0.455	Mean= 0.682 Sd= 0.410
loglik	-5274	-5274	-5441	-5310
AIC	10552	10552	10885	10625
BIC	10556	10556	10900	10639
p-value	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$

Табл. 2: Гамма-розподіл

	Методи одержання оцінок			
	mme	mle	qme	mge
loglik	-4821	-4725	-4869	-4743
AIC	9646	9454	9741	9489
BIC	9660	9468	9756	9509
p-value	$3.323 \cdot 10^{-9}$	$7.064 \cdot 10^{-12}$	$2.964 \cdot 10^{-14}$	$5.142 \cdot 10^{-5}$

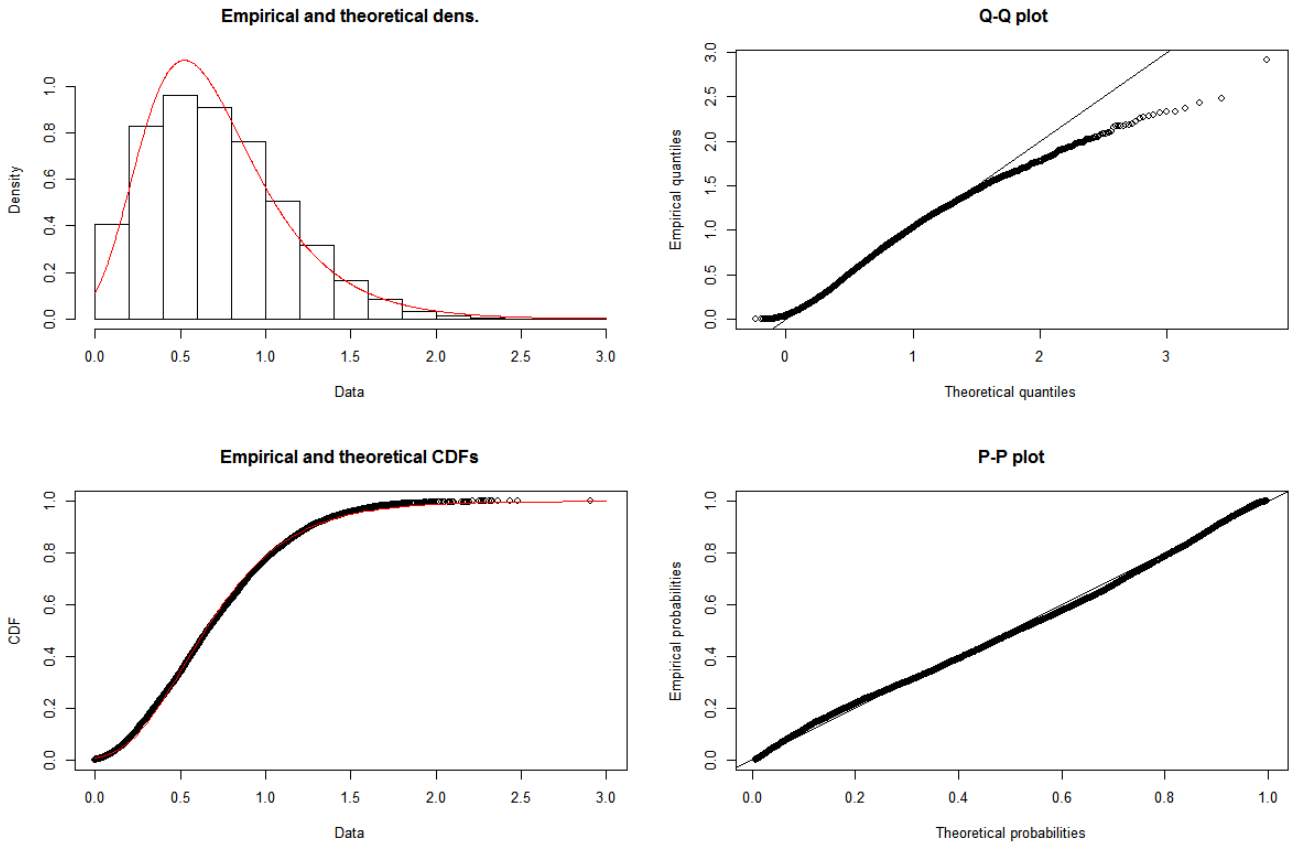

Рис. 5: Нормальний розподіл

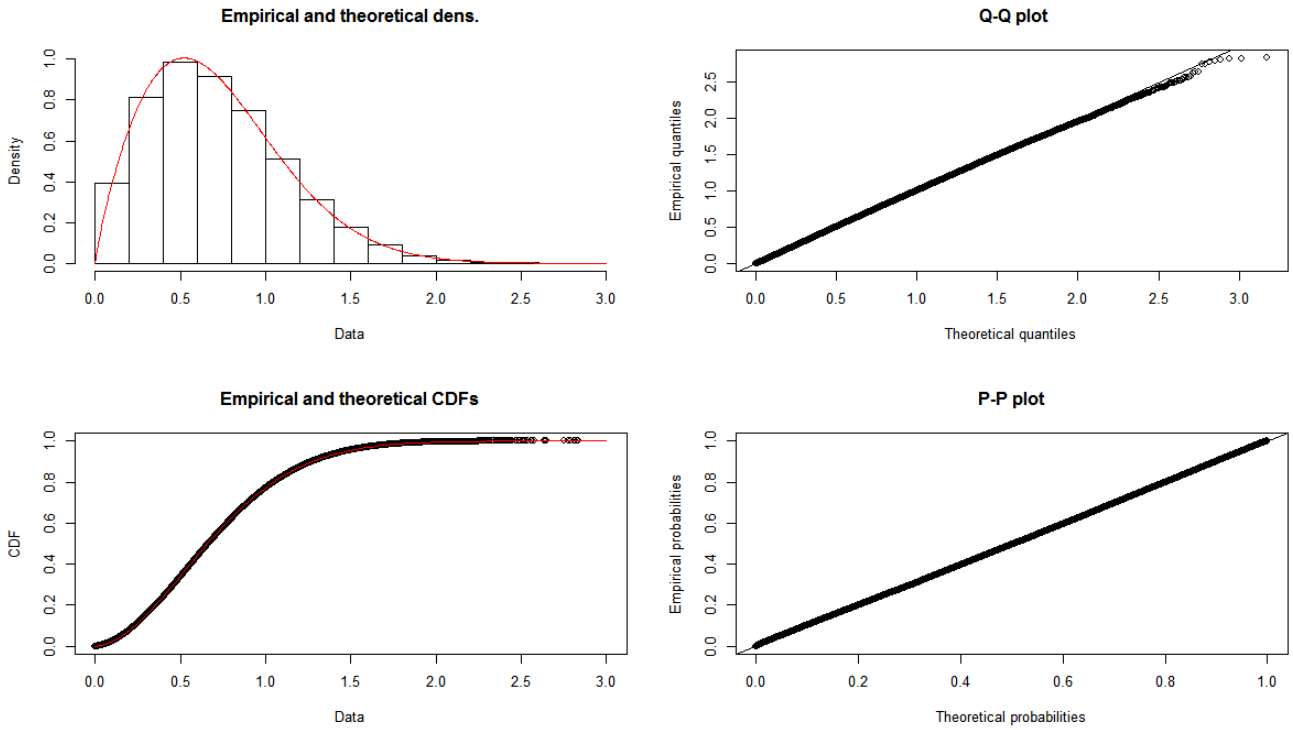

Рис. 6: Нормальний розподіл

гіпотез функцією `mledist`.

Табл. 3: Порівняння розподілів Ґумбеля та Вейбула

	Розподіл		
	Логнормальний	Ґумбеля	Вейбула
Оцінки	Meanlog= -0.539 Sdlog= 0.730	A= 0.525 B= 0.333	Shape= 1.804 Scale= 0.8065
loglik	-5641	-4757	-4547
p-value	$< 2.2 \cdot 10^{-16}$	$9.664 \cdot 10^{-7}$	0.4475

Наведемо фрагмент коду програми.

```
fit.norm <- fitdist(data = m, "norm", method = "mge", gof="CvM")
print(fit.norm)
plot(fit.norm)
ks.test(unique(m),"pnorm",mean=fit.norm$estimate[1],sd= fit.norm$estimate[2])
fit.gama <- fitdist(data = m, "gamma", method = "mme")
print(fit.gama)
plot(fit.gama)
ks.test(unique(m),"pgamma",shape=fit.gama$estimate[1],rate=fit.gama$estimate[2])
fit.exp <- fitdist(data = m, "exp", method = "mme")
print(fit.exp)
plot(fit.exp)
ks.test(unique(m),"pexp",rate=fit.exp$estimate[1])
fit.logis <- fitdist(data = m, "logis", method = "mme")
print(fit.logis)
plot(fit.logis)
ks.test(unique(m),"plogis",location=fit.logis$estimate[1],scale=fit.logis$estimate[2])
f3<-mledist(m[m>0],"weibull",lower = c(0, 0))
```

```
f3
plotdist(m[m>0], "weibull", para=list(shape=f3$estimate[1], scale=f3$estimate[2]))
ks.test(unique(m[m>0]),"pweibull",shape=f3$estimate[1],scale=f3$estimate[2])
dgumbel <- function(x,a,b) 1/b*exp((a-x)/b)*exp(-exp((a-x)/b))
pgumbel <- function(q,a,b) exp(-exp((a-q)/b))
qgumbel <- function(p,a,b) a-b*log(-log(p))
f4<-mledist(m,"gumbel",start=list(a=0,b=2),optim.method="Nelder-Mead")
f4
ks.test(unique(m),"pgumbel",a=f4$estimate[1],b=f4$estimate[2])
plotdist(m, "gumbel", para=list(a=f4$estimate[1], b=f4$estimate[2]))
```

3 Висновки

Змодельовано за новим алгоритмом вибірку, що імітує розподіл максимуму звуження двопараметричного поля Ченцова на певну криву.

Проведено підбір закону розподілу вибірки.

Знайдено оцінки параметрів щільностей розподілу, одержаних методами: моментів, квантілів, максимальної вірогідності, найменшої відстані.

Перевірено гіпотезу про узгодженість законів розподілів за допомогою критерію Колмогорова–Смірнова, а також критеріїв AIC, BIC.

Показано, що оцінки, одержані різними методами, відрізняються між собою незначним чином, крім методу квантілів. Оцінки, одержані цим методом, дали найбільшу похибку. Найкраще значення p-value для критерія Колмогорова–Смірнова маємо для методу найменшої відстані. Але це очевидно, з огляду вибраних метрик, оскільки вони фактично повторюють статистики критеріїв узгодженості.

Аналіз проводився за допомогою мови статистичного програмування R. Можна зробити висновок, що середовище програмування R — це швидкий і зручний засіб для статистичного аналізу вибірок, а пакет `fitdistrplus` безумовно ефективним у вирішенні поставленої задачі.

References

Chentsov, N. N. (1956). Wiener random fields depending on several parameters. *Doklady Akademii Nauk SSSR*, *106*(4), 607–609.

The Comprehensive R Archive Network. (n.d.). https://cran.cnr.berkeley.edu/

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton, NJ: Princeton University Press.

Dykhovychnyi, O. O., & Kruglova, N. V. (2018). Simulation of a gaussian process with correlation function of a special form. In *Abstracts of International conference "Stochastic Equations, Limit Theorems and Statistics of Stochastic Processes dedicated to the 100th anniversary of I. I. Gikhman, 2018, September 17–22, Kyiv,*

Ukraine (pp. 18–19).
http://matan.kpi.ua/gikhman100conf/g100-abstracts.pdf

Kobzar, A. I. (2006). *Applied mathematical statistics.* Moscow: Fizmatlit.

Lapko, A. V., Chentsov, S. V., Krokhov, S. I., & Feldman, L. A. (1996). *Self-learning systems for information processing and decision making.* Novosibirsk: Nauka.

Park, C., & Paranjape, S. R. (1974). Probabilities of Wiener paths crossing differentiable curves. *Pacific journal of mathematics*, *53*(2), 579–583.
https://projecteuclid.org/euclid.pjm/1102911625

Syzrantsev, V. N., Nevelev, Y. P., & Golofast, S. L. (2006). Adaptive method for probability density function reconstruction. *Proceedings of Higher Educational Institutions. Machine Building*, *2006*(12), 3–11.

Wolfowitz, J. (1957). The minimum distance method. *The Annals of Mathematical Statistics*, *28*(1), 75–88.
https://doi.org/10.1214/aoms/1177707038

Wolverton, C. T., & Wagner, T. J. (1969). Asymptotically optimal discriminant functions for pattern classification. *IEEE Transactions on Information Theory*, *15*(2), 258–265.
https://doi.org/10.1109/TIT.1969.1054295

Yeh, J. (1960). Wiener measure in a space of functions of two variables. *Transactions of the American Mathematical Society*, *95*(3), 433–450.
https://doi.org/10.1090/S0002-9947-1960-0125433-1

О. О. Диховичний, Н. В. Круглова, О. І. Вирстюк (2018). Методи знаходження законів розподілів випадкових величин за даними статистичних вибірок засобами мови R. *Mathematics in Modern Technical University*, 2018(1), 91–100.

Submitted: 2018-10-01
Accepted: 2018-11-14

O. O. Dykhovychnyi, N. V. Kruglova, O. I. Virstiuk (2018). Methods for identification of probability distribution of random variables from data samples with R statistical computing language. *Mathematics in Modern Technical University*, 2018(1), 91–100.

Abstract. The following article discusses various methods for probability distribution fitting to simulated data by means of R statistical computing language. In particular, some graphical methods like plotting of histograms, empirical and theoretical density functions, P-P and Q-Q plots, were considered. Estimation functions for probability distribution parameters were investigated by applying method of moments, method of quantiles, method of maximum likelihood, and shortest distance method. Hypothesis about probability distribution were verified with Kolmogorov–Smirnov, AIC, and BIC tests.

The corresponding data set used to illustrate the above methods was taken from probability distribution of the maximum of Chenstov field restriction to a particular curve. The distribution was simulated with the special original algorithm in R statistical software.

Keywords: R language; Chentsov field; Gaussian process; method of moments; method of quantiles; maximum likelihood estimation; minimum distance estimation; statistical tests.