УДК 519.2 : 519.7

В.Ю. Дубницкий, Л.Д. Филатова, А.И. Ходырев

Харьковский учебно-научный институт ГВУЗ «Университет банковского дела», Харьков

УСТОЙЧИВОСТЬ ОЦЕНКИ ЭНТРОПИИ ГИСТОГРАММЫ НЕПРЕРЫВНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ПО ОТНОШЕНИЮ К ИЗМЕНЕНИЮ КОЛИЧЕСТВА ЕЁ ИНТЕРВАЛОВ

Разработаны предложения к методике определения устойчивости оценки энтропии гистограммы непрерывной случайной величины по отношению к изменению количества её интервалов. Предложен способ определения оценки устойчивости оценки энтропии гистограммы непрерывной случайной величины по отношению к изменению количества её интервалов. Проведен численный эксперимент, в процессе которого для различных способов определения количества интервалов гистограммы вычислено среднее значение энтропии и её верхняя и нижняя доверительные границы. Для нормального, логистического, гамма-распределения и распределения Вейбулла по результатам непараметрического дисперсионного анализа установлено, что оценка энтропии непрерывной случайной величины неустойчива к виду закона распределения, количеству интервалов гистограммы и, следовательно, к количеству наблюдений.

***Ключевые слова**: энтропия, гистограмма, оценка энтропии, непараметрический дисперсионный анализ, нормальное распределение, логистическое распределение, гамма- распределение, распределение Вейбулла.*

Введение

Понятие энтропии хорошо известно и широко используется при решении задач, связанных с теорией передачи сигналов в условиях помех в каналах связи. В соответствии с работой [1] энтропией Шеннона непрерывной случайной величины X, имеющей плотность распределения f(x), называют функционал вида:

$$H(X) = -\int_{-\infty}^{\infty} f(x)\log_a f(x)dx \text{ (ед)}. \qquad (1)$$

Если изучаемая система может пребывать в одном из m возможных состояний, то её энтропию определяют по условию:

$$H(X) = -\sum_{i=1}^{m} p_i \log_a p_m \text{ (ед)}; \qquad (2)$$

где p_i – вероятность пребывания системы в одном из i (i=1,2…m) возможных состояний. Если основание логарифма α равно 10, то единицу энтропии называют дит, если величина α равна основанию натурального логарифма, то единицу энтропии называют нит, если величина α=2, то единицу энтропии называют бит. Соотношения между этими величинами приведены в табл. 1. Данные, приведенные во втором столбце табл. 1 заимствованы из работы [4].

В данном сообщении единицей измерения энтропии принят нит. Вычисление величины H(x) для различных законов распределения подробно рассмотрено в работах [2, 3]. По аналогии с взаимоотношениями между теорией вероятности и математической статистикой можно говорить, что энтропия, вычисленная по условиям (1, 2), вычислена при известном законе распределения и известных значениях его параметров.

Таблица 1

Соотношения между единицами измерения энтропии

$\lg N = 0.4343 \ln N$	1*дит* = 2,3 *нит*
$\lg N = 0.3010 \log_2 N$	1 *дит* = 3,3 *бит*
$\ln N = 2.3025 \lg N$	1 *нит* = 0,43 *дит*
$\ln N = 0.6931 \log_2 N$	1 *нит* =1,45 *бит*
$\log_2 N = 3.3225 \lg N$	1 *бит* =0,3 *дит*
$\log_2 N = 1.4427 \ln N$	1 *бит* = 0,69 *нит*

Если закон распределения случайной величины определён по данным, полученным по некоторой выборке, то определённую на основе этого величину энтропии h(X) будем называть оценкой величины H(x). Так как чаще всего источником сведений о свойствах случайных величин служат их гистограммы, построенные по результатам выборочных данных, то в рамках данной работы будет рассмотрено влияние характеристик гистограммы на оценку энтропии, величину h(X).

Анализ литературы

В работе [4] предложено для оценки энтропии случайной величины, заданной своей гистограммой, использовать выражение вида:

$$h(x) = \sum_{i=1}^{m} \frac{n_i}{n} \ln \frac{n}{n_i} + \ln d\,. \qquad (3)$$

В условии (3) принято, что все интервалы гистограммы имеют равный шаг d. В этой же работе приведены выражения, позволяющие оценить статистические свойства величины h(X). В работе [5] дисперсию D_h оценки энтропии h(X) предложено

определять по условию:

$$D_h = \frac{1}{n}\left[\sum_{i=1}^{m} \hat{p}_i \ln^2 \hat{p}_i - \left(\hat{p}_i \ln \hat{p}_i\right)^2\right] + \frac{m-1}{2n^2}; \quad (4)$$

где n – количество наблюдений (объём выборки), m – количество интервалов гистограммы, $\hat{p}_i$ – частота i-ого интервала гистограммы, $\hat{p}_i = n_i / n$; n_i = количеству наблюдений, попавших в i-ый интервал гистограммы. В этой же работе доверительный интервал гистограммы предлагают определять по условию:

$$I_h = \left[h - u\left(\frac{1+\beta}{2}\right)\frac{D_h}{\sqrt{m}}; h - u\left(\frac{1+\beta}{2}\right)\frac{D_h}{\sqrt{m}}\right]; \quad (5)$$

где $u(\cdot)$ – квантиль стандартного нормального распределения. При выполнении расчётов в данном сообщении эта величина принята равной 1,64. Из рассмотрения условий (2…5) следует, что важнейшим параметром, который может влиять на полученный результат является количество интервалов гистограммы m. Способы определения этой величины как функции объёма выборки n подробно рассмотрены в работе [6]. В данной работе принят объём выборки n=200. При выполнении расчетов использованы следующие выражения:

$$m = 3{,}3\lg(n) + 1; \quad (6)$$

$$m = 5\lg(n); \quad (7)$$

$$m = \sqrt{n}; \quad (8)$$

$$m = 4\sqrt[5]{2} \cdot \left(n / u(\cdot)\right)^{0,4}; \quad (9)$$

$$m = 4\left[0{,}75\left(n-1\right)^2\right]^{0,2}; \quad (10)$$

$$m = 1{,}9n^{0,4}. \quad (11)$$

Результаты определения количества интервалов гистограммы для всех приведенных в работе [6] выражений приведены в табл. 2.

Таблица 2

Определение количества интервалов гистограммы

Расчётная формула	Ф(6)	Ф(7)	Ф(8)	Ф(9)	Ф(10)	Ф(11)
Количество интервалов	9	11	14	31	31	16

Для принятого в работе объёма выборки количество интервалов m=31 отвергнуто, как лишённое содержательного смысла при принятом объёме выборки.

В работах [7…10] показано, что оценки вида (3), (4) имеют асимптотически нормальное распределение. Установление этого факта позволило получить оценку для доверительного интервала энтропии гистограммы вида (6) и получены оценки асимптотического смещения оценки энтропии и её дисперсии. В указанных работах сформулированы достаточные условия получения асимптотически нормальной оценки энтропии. То есть, обоснованы основные теоретические способы получения оценки энтропии. Методическая особенность этих работ в том, что количество интервалов гистограммы заранее определено и влияние изменения этой характеристики на оценки энтропии не рассматривались. Исторический обзор развития метода гистограмм приведен в работе [11]. Современное состояние задачи о построении гистограмм рассмотрено в работе [12]. Влияние количества интервалов гистограммы на решение задачи о выборе закона распределения случайной величины рассмотрено в работе [13, 14]. В работе [15] показано применение понятия энтропии гистограммы для обоснования закона распределения, характерного для данной гистограммы. Таким образом, из проведенного обзора литературы следует, что мостик, связывающий влияние количества интервалов гистограммы на её энтропию, отсутствует.

Постановка задачи: разработка предложений к методике определения устойчивости оценки энтропии гистограммы непрерывной случайной величины по отношению к изменению количества её интервалов.

Полученные результаты

Для получения исходных данных был проведен численный эксперимент. Для его проведения было принято, что все генерируемые псевдослучайные выборки, предназначенные для последующего анализа, получали при условии, что их среднее значение m = 200 и среднеквадратическое отклонение s = 40.

Генерируемые выборки, в нашем случае, должны были соответствовать нормальному распределению, логистическому распределению, гамма-распределению и распределению Вейбулла. Сведения об основных законах распределения случайных величин, использованных в работе, приведены в табл. 3. В этой таблице и далее принято, что греческие буквы соответствуют параметрам функций плотности, латинские буквы – их основным числовым характеристикам: m – математическому ожиданию, s – среднеквадратическому отклонению, v – коэффициенту вариации. Для определения параметров принятых в работе законов распределения использованы результаты работ [15, 16]. Сведения о параметрах законов распределения, принятых в качестве исходных при моделировании, приведены в табл. 4.

Для каждого из указанных в табл. 3, 4 законов распределения генерировали по 200 псевдослучайных чисел. Этот процесс выполняли, используя систему STATGRAPHICS V.1. Результаты моделирования приведены в табл. 5, 6. В табл. 5 приведены основные статистические характеристики выборок, полученных в результате моделирования соответствующих распределений. В табл. 6 приведены параметры законов распределения, определённые для выборок, полученных в результате моделирования соответствующих распределений.

Таблица 3

Основные характеристики функций плотности распределения вероятности

Функции плотности распределения при $-\infty < x < \infty$		
Тип распределения	Плотность распределения	Зависимость параметров распределения от его начальных характеристик
Нормальное распределение	$f(x) = \left(\sigma\sqrt{2\pi}\right)^{-1} \cdot \exp\left(-(x-\mu)^2 \Big/ \left(2\sigma^2\right)\right)$	$\mu = m;\ \sigma = s$
Логистическое распределение	$f(x) = \dfrac{\exp\left((x-\mu)/\lambda\right)}{\lambda\left[1+\exp\left((x-\mu)/\lambda\right)\right]^2} = \dfrac{1}{4\lambda \mathrm{ch}^2\left((x-\mu)/2\lambda\right)}$	$\mu = m;\ \lambda = \dfrac{s\sqrt{3}}{\pi} = 0{,}55133s$
Функции плотности распределения при $0 \le x < \infty$		
Гамма-распределение	$f(x) = \left(\lambda^{\alpha} \Big/ \Gamma(\alpha)\right) \cdot x^{\alpha-1} e^{-\lambda x},\ x > 0$	$\alpha = 1/v^2;\ \lambda = m/s^2$
Распределение Вейбулла	$f(x) = (c/\alpha)(x/\alpha)^{c-1} \exp\left[-(x/\alpha)^c\right]$	Выражение в явном виде отсутствует

Таблица 4

Начальные значения параметров законов распределения вероятности, принятые при моделировании

Типы законов распределения, использованные в работе							
Нормальный		Логистический		Гамма-распределение		Вейбулла	
μ	λ	μ	λ	α	λ	α	c
200	40	200	22,052	25	0,125	215.93098	5,82824

Таблица 5

Средние значения и среднеквадратические отклонения полученных в результате моделирования выборок

Типы законов распределения, использованные в работе							
Нормальный		Логистический		Гамма-распределение		Вейбулла	
m	s	m	s	m	s	m	s
197,5519	39,1722	197,4389	44,4990	198,827	42,1772	203,891	40,7029

Таблица 6

Значения параметров законов распределения вероятности, полученные при моделировании

Типы законов распределения, использованные в работе							
Нормальный		Логистический		Гамма-распределение		Вейбулла	
μ	λ	μ	λ	α	λ	α	c
197,5519	39,1722	197,4389	21,5956	25,40485	0,11768	220,125	5,825

Сравнение результатов, приведенных в табл. 2 и табл. 6 позволяет сделать вывод о том, что параметры законов распределения для полученных выборок вполне удовлетворительно соответствуют первоначально заданным значениям и могут быть использованы для дальнейшего анализа. Используя программную систему Atte Stat, реализующую процесс вычислений условий (3...5) для всех указанных выше законов распределений, определены нижнее доверительное на уровне 0,95 значение энтропии $h^{-}(x)$, среднее значение энтропии h(x) и её верхнее доверительное значение $h^{+}(x)$. Все вычисления проведены для интервалов, количество которых указано в табл. 2. Результаты вычислений приведены в табл. 7.

Таблица 7

Оценка энтропии для выбранных законов распределения

Вид закона распределения	Количество интервалов	Оценка энтропии (нит)		
		Нижнее доверительное значение $h^{-}(x)$	Среднее значение $h^{-}(x)$	Верхнее доверительное значение $h^{+}(x)$
Нормальный	9	1,951337	1,978717	2,006123
	11	2,14735	2,172424	2,197490
	14	2,364088	2,386382	2,408677
	16	2,512838	2,533772	2,554706
Логистический	9	1,488867	1,526402	1,563938
	11	1,680299	1,713566	1,746833
	14	1,935790	1,962951	1,990115
	16	2,058687	2,083275	2,114287
Гамма-распределение	9	2,044215	2,065975	2,087739
	11	2,209686	2,232114	2,254535
	14	2,451659	2,472181	2,492703
	16	2,571364	2,595528	2,611697
Вейбулла	9	1,801036	1,832229	1,863405
	11	2,209686	2,232114	2,254535
	14	2,451659	2,472181	2,4927603
	16	2,571365	2,591528	2,611697

На этом этапе подготовку данных для решения поставленной задачи – определение устойчивости оценки энтропии гистограммы непрерывной случайной величины по отношению к изменению количества её интервалов можно считать завершенной. Рабочая гипотеза, проверяемая в данной работе следующая: вид закона распределения и количество интервалов разбиения существенно влияют на среднее значение энтропии, полученное по заданной гистограмме. Для этого введём коэффициент:

$$\eta_{uk} = h_{uk}(x)/h_{1k}(x). \qquad (12)$$

где u – количество интервалов гистограммы, u = 1…4 (1 – 9 интервалов, 2 – 11 интервалов, 3 – 14 интервалов, 4 – 16 интервалов гистограммы); k – вид закона распределения выборки, для которой построена гистограмма, k = 1…4 (1 – нормальный закон распределения, 2 – логистический, 3 – гамма-распределение, 4 – распределение Вейбулла). Для большей наглядности это отношение приведено в децибелах (dB) и показателе «Отличие в разы».

Обоснование такого способа измерения отношения двух величин дано в работах [17, 18]. Результаты соответствующих вычислений приведены в табл. 8. График, иллюстрирующий это влияние для показателя «отличие в разы», показан на рис. 1. На рисунке значение показателя «отличие в разы» представлено на оси ординат, на оси абсцисс показано количество интервалов, вид закона распределения ясен из принятых условных обозначений. Из графиков, показанных на рисунке можно предположить, что исследуемые факторы влияют на устойчивость полученной оценки. Для статистического обоснования сделанного предположения использован двухфакторный непараметрический дисперсионный анализ. Исходные данные для анализа показателя «отличия в разы» приведены в табл. 9.

Таблица 8

Оценка влияния количества интервалов гистограммы и вида закона распределения на величину коэффициента η_{uk}

Вид распределения	Количество интервалов гистограммы (m)	Среднее значение энтропии (нит)	Децибелы, dB	Отличие в разы
Нормальное	9	1,978717	0	1
	11	2,172424	0,405609	1,1
	14	2,386382	0,813563	1,21
	16	2,533772	1,073838	1,28
Логистическое	9	1,526402	0	1
	11	1,713566	0,502319	1,12
	14	1,962951	1,092405	1,29
	16	2,083275	1,350777	1,36
Гамма-распределение	9	2,065975	0	1
	11	2,232114	0,335913	1,08
	14	2,472181	0,779552	1,2
	16	2,595528	0,991007	1,26
Вейбулла	9	1,832229	0	1
	11	2,232114	0,857366	1,22
	14	2,472181	1,301005	1,35
	16	2,591528	1,505762	1,41

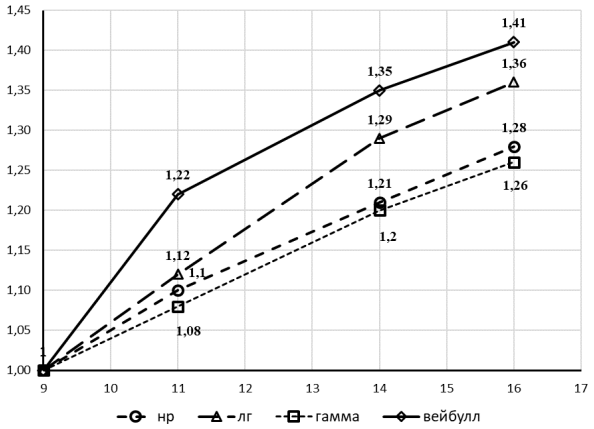

Рис. 1. Изменение показателя «отличие в разы» в зависимости от вида закона распределения и количества интервалов гистограммы

Таблица 9

Исходные данные для двухфакторного дисперсионного анализа показателя «отличия в разы»

Количество интервалов гистограммы	Вид закона распределения			
	Нормальный	Логистический	Гамма-распределение	Вейбулла
9	1,00	1,00	1,00	1,00
11	1,10	1,12	1,08	1,22
14	1,21	1,29	1,20	1,35
16	1,28	1,36	1,26	1,41

Для анализа этих данных использованы критерии Фридмана и Пэйджа, основы применения которых описаны в работе [5]. Критерий Фридмана проверяет нулевую гипотезу о том, что отсутствует эффект столбца. В нашем случае это означает, что изменение закона распределения гистограммы не влияет на устойчивость оценки энтропии. Критерий Пейджа проверяет нулевую гипотезу о том, что отсутствует эффект столбца и строки. В нашем случае это означает, что все рассмотренные в работе факторы не влияют на оценку устойчивости энтропии, рассматриваемую в данной работе. В результате вычислений установлено, что для критерия Фридмана величина Pv = 0,06. Для критерия Пейджа величина Pv = 0,30. Эти величины превосходят критическую величин Pv = 0,05. Таким образом, следует в каждом случае принять альтернативную гипотезу о том, что оценка энтропии непрерывной случайной величины неустойчива к виду закона распределения, количеству интервалов гистограммы и, как это следует из условий (6…11) к количеству наблюдений.

Выводы

1. Разработаны предложения к методике определения устойчивости оценки энтропии гистограммы непрерывной случайной величины по отношению к изменению количества её интервалов.

2. Предложен способ определения оценки устойчивости оценки энтропии гистограммы непрерывной случайной величины по отношению к изменению количества её интервалов.

3. Проведен численный эксперимент, в процессе которого для различных способов определения количества интервалов гистограммы вычислено среднее значение энтропии и её верхняя и нижняя доверительные границы.

4. Для нормального, логистического, гамма-распределения и распределения Вейбулла по результатам непараметрического дисперсионного анализа установлено, что оценка энтропии непрерывной случайной величины неустойчива к виду закона распределения, количеству интервалов гистограммы и, следовательно, к количеству наблюдений.

Список литературы

1. Кузьмин И.В. Основы теории информации и кодирования / И.В. Кузьмин, В.А. Кедрус. – К.: Вища школа, 1986. – 238 с.

2. Заездный А.М. Основы расчетов по статистической радиотехнике / А.М. Заездный – М.: Связь, 1969. – 447 с.

3. Michlowicz J. V. Handbook of DIFFERENTIAL ENTROPY / J.V. Michlowicz, J.M. Nichols, Bucholtz F. – New York.: A.CHAPMAN & HALL, 2014. – 220 p.

4. Электрические измерения неэлектрических величин / [А.М. Туричин, П.В. Новицкий, Е.С. Левшина и др.] под ред.П.В. Новицкого. – Л-д.: Энергия, 1975. – 576 с.

5. Гайдышев И.П. Моделирование стохастических и детерминированных систем: Руководство пользователя программы Atte Stat / И.П. Гайдышев. БИ, 2015. – 484 с.

6. Новицкий П.В. Оценка погрешностей результатов измерений / П.В. Новицкий, И.А. Зограф. – Л-д.: Энергоатомиздат, 1991. – 304 с.

7. Башарин Г.П. О статистической оценке энтропии независимых случайных величин / Г.П. Башарин // Теория вероятностей и её применение. – 1959. – Т.IV, №3. – С. 361-364.

8. Добрушин Р.Л. Упрощенный метод экспериментальной оценки энтропии случайных последовательностей / Р.Л. Добоушин // Теория вероятностей и её применение. – 1958. – Т.3, вып. 4. – С. 462-464.

9. Зубков А.М. Предельные распределения статистической оценки энтропии / А.М. Зубков // Теория вероятностей и её применение. – 1973. – Т. 18, вып. 3. – С. 643-650.

10. Михайлов В.Г. Статистическое оценивание энтропии дискретных случайных величин с большим числом исходов / В.Г. Михайлов, В.А. Ватутин // Успехи математических наук. – 1995. – Т. 50, вып. 5 (305). – С. 121-134.

11. Ioannidis Y. The history of histograms / Y. Ioannidis // Proceeding 2003 VL Conference/ 2003. pp 19-30.

12. Битюков С.И. Сравнение гистограмм в физических исследованиях / С.И. Битюков, А.В. Максимушкина, В.В. Смирнова Сравнение гистограмм в физических исследованиях // Изв.ВУЗов, сер. «Ядерная физика». – 2016. – №1. – С. 81-90.

13. Петрушин В.Н. Бикритериальный метод построения и оценки качества гистограмм / В.Н. Петрушин, М.В. Ульянов, И.А. Чертихина, Е.В. Никульчев // Информационные технологии и вычислительные системы. – 2012. – №4 – С. 3-12.

14. Тарасов И.Е. О выборе интервалов гистограммирования / И.Е. Тарасов // Системы управления и информационные технологии. – 2011. – № 2.1 (44) – С. 181-184.

15. Вадзинский Р.Н. Справочник по вероятностным распределениям. / Р.Н. Вадзинский. – М.: НАУКА, 2001. – 295 с.

16. Дубницкий В.Ю. Оптимальная аппроксимация функции плотности распределения информации по критерию минимума потери информации / В.Ю. Дубницкий, И.Г. Скорикова, А.И. Ходырев // Системи обробки інформації. – Х.: ХНУПС, 2017. – Вип. 4. – С. 45-51.

17. Зельдин Е.А. Децибелы / Е.А. Зельдин. – М.: Энергия, 1977. – 64 с.

18. Дубницкий В.Ю. Определение относительной оценки тяжести хвоста распределения – уровня хвоста / В.Ю. Дубницкий, А.И. Ходырев // Системи обробки інформації. – Х.: ХУПС, 2015. – Вип. 7 (132). – С. 83-92.

Надійшла до редколегії 1.08.2017

Рецензент: д-р техн. наук, проф. О.О. Можаєв, Національний технічний університет «ХПІ», Харків.

СТІЙКІСТЬ ОЦІНКИ ЕНТРОПІЇ ГІСТОГРАМИ НЕПЕРЕРВНОЇ ВИПАДКОВОЇ ВЕЛИЧИНИ ПО ВІДНОШЕННЮ ДО ЗМІНИ КІЛЬКОСТІ ЇЇ ІНТЕРВАЛІВ

В.Ю. Дубницький, Л.Д. Філатова, О.І. Ходирєв

Розроблено пропозиції до методики визначення стійкості оцінки ентропії гістограми неперервної випадкової величини по відношенню до зміни кількості її інтервалів. Запропоновано спосіб визначення оцінки стійкості оцінки ентропії гістограми безперервної випадкової величини по відношенню до зміни кількості її інтервалів. Проведений чисельний експеримент в процесі якого для різних способів визначення кількості інтервалів гістограми обчислено середнє значення ентропії і її верхня і нижня довірчі межі. Для нормального, логістичного, гамма- розподілу та розподілу Вейбулла за результатами непараметричного дисперсійного аналізу встановлено, що оцінка ентропії неперервної випадкової величини нестійка до виду закону розподілу, кількості інтервалів гістограми та до кількості спостережень.

***Ключові слова**: ентропія, гістограма, оцінка ентропії, непараметричний дисперсійний аналіз, нормальний розподіл, логістичний розподіл, гамма-розподіл, розподіл Вейбулла.*

EVALUATION STEADFASTNESS OF A CONTINUOUS RANDOM QUANTITY HISTOGRAM ENTHROPY RELATIVE TO ALTERNATING NUMBER OF ITS INTERVALS

V.Yu. Dubnitskiy, L.D. Filatova, A.I. Khodyrev

Proposals were developed for evaluation steadfastness determination method of continuous random quantity histogram entropy relative to alternating number of its intervals. A determination process was proposed for evaluation steadfastness of continuous random quantity histogram entropy relative to alternating number of its intervals. Numerical experiment was performed and in its course average value of entropy and its top and bottom confidence limits for various determination methods of number of histogram intervals. It was found for normal distribution, logistical distribution, gamma distribution and Weibull's distribution by non-parametric dispersion analysis, that continuous random quantity entropy evaluation is not steadfast to the type of distribution law, to histogram number of intervals and to the number of observations.

***Keywords:** entropy, histogram, entropy evaluation, non-parametric dispersion analysis, normal distribution, logistical distribution, gamma distribution, Weibull's distribution.*