

# ГРАФОВІ ЙМОВІРНІСНІ МОДЕЛІ СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ Data Mining

Сумський державний університет

# Основні визначення

Ймовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Математична статистика — це прикладна математична дисципліна, яка примикає до теорії ймовірностей. Вона базується на поняттях і методах теорії ймовірностей, але вирішує свої специфічні завдання спеціальними методами.

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Математична статистика — це прикладна математична дисципліна, яка примикає до теорії ймовірностей. Вона базується на поняттях і методах теорії ймовірностей, але вирішує свої специфічні завдання спеціальними методами.
- Основне завдання математичної статистики — отримати обгрунтовані висновки про параметри, видах розподілів та інших властивостях випадкових величин по кінцевій сукупності спостережень над ними.

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Ми зупинимося на основних методах аналізу **одновимірних** статистичних даних: визначення точкових та інтервальних оцінок параметрів розподілу, перевірка гіпотез про вид розподілу.

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Ми зупинимося на основних методах аналізу **одновимірних** статистичних даних: визначення точкових та інтервальних оцінок параметрів розподілу, перевірка гіпотез про вид розподілу.
- Також ознайомимося з елементами кореляційного, дисперсійного і регресійного аналізу **двовимірних** статистичних даних.

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Генеральною сукупністю називають всю сукупність реалізації випадкової величини  $X$ , всі можливі спостереження деякого показника, всі можливі результати деякого випробування.

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Генеральною сукупністю називають всю сукупність реалізації випадкової величини  $X$ , всі можливі спостереження деякого показника, всі можливі результати деякого випробування.
- Вибіркою називають частину генеральної сукупності  $X_n = \{x_1, x_2, \dots, x_n\}$ , тобто кінцеве підмножина значень випадкової величини з безлічі елементів генеральної сукупності.

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Генеральною сукупністю називають всю сукупність реалізації випадкової величини  $X$ , всі можливі спостереження деякого показника, всі можливі результати деякого випробування.
- Вибіркою називають частину генеральної сукупності  $X_n = \{x_1, x_2, \dots, x_n\}$ , тобто кінцеве підмножина значень випадкової величини з безлічі елементів генеральної сукупності.
- Об'ємом вибірки  $n$  називають кількість випадкових величини  $X$ , що в ній містяться.



# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

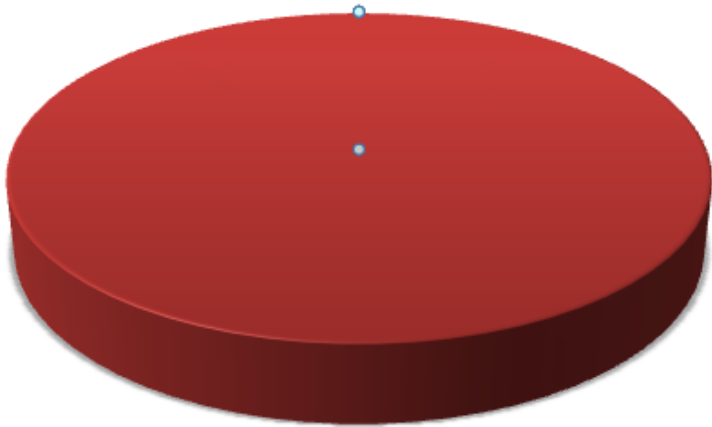
Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

## Генеральна сукупність



# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

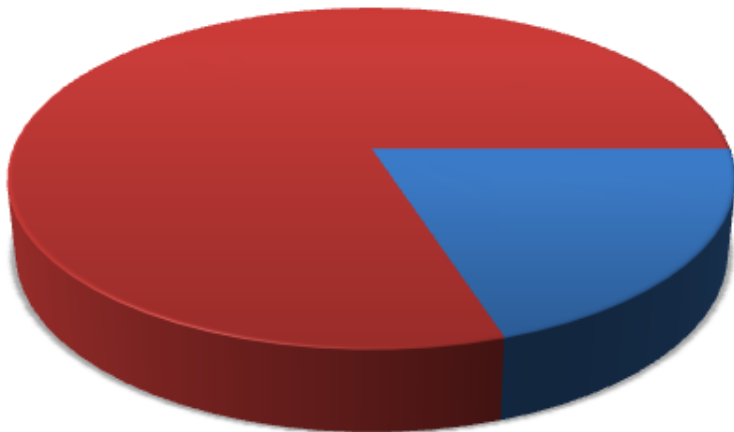
Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Генеральна сукупність + вибірка



# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

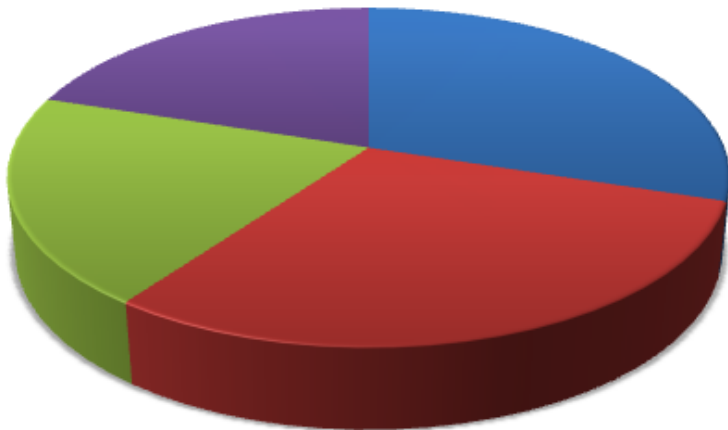
Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Генеральна сукупність + вибірка



# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

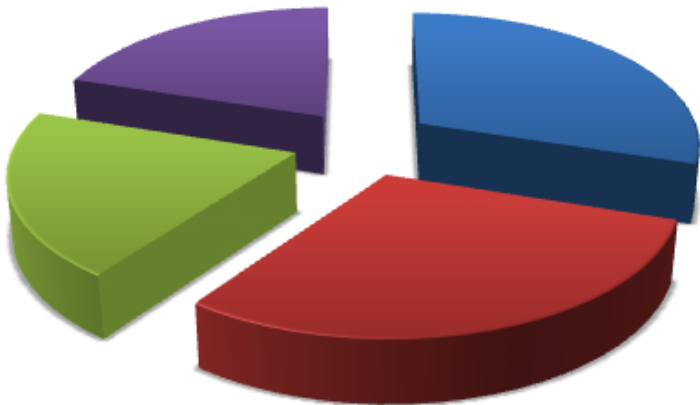
Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

## Вибірка



# Основні визначення

## Вибірка



Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Завдання математичної статистики полягає в дослідженні властивостей вибірки та узагальненні цих властивостей на всю генеральну сукупність.

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Завдання математичної статистики полягає в дослідженні властивостей вибірки та узагальненні цих властивостей на всю генеральну сукупність.
- Вибірка є вихідною інформацією для статистичного аналізу та прийняття рішень про невідомі імовірнісні характеристики випадкової величини  $X$ .

# Основні визначення

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Завдання математичної статистики полягає в дослідженні властивостей вибірки та узагальненні цих властивостей на всю генеральну сукупність.
- Вибірка є вихідною інформацією для статистичного аналізу та прийняття рішень про невідомі імовірнісні характеристики випадкової величини  $X$ .
- Для того щоб за вибіркою можна було досить впевнено судити про генеральну сукупність, вибірка повинна бути **репрезентативною**, тобто досить повно представляти ознаки і параметри генеральної сукупності. Репрезентативність вибірки поліпшується при збільшенні її об'єму.



# Основні визначення

## Вибірка



Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

# Задачі статистичного аналізу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



# Задачі статистичного аналізу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



# Задачі статистичного аналізу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



# Задачі статистичного аналізу

Імовірнісні  
основи  
обробки  
даних

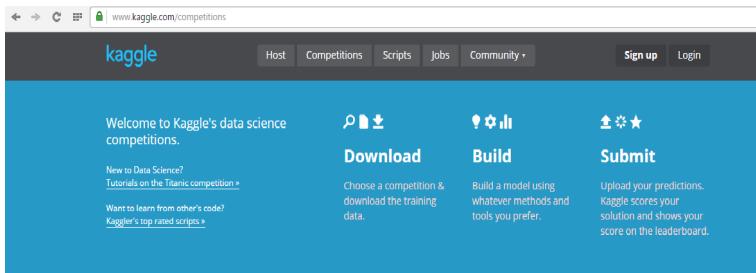
Основні  
визначення

Задачі ста-  
тистичного  
аналізу



Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



The screenshot shows the Kaggle website's main interface. At the top, there's a navigation bar with the Kaggle logo and links for Host, Competitions, Scripts, Jobs, and Community. On the right, there are buttons for Sign up and Login. The main content area is divided into four columns: 1. Welcome to Kaggle's data science competitions, with links for new users and learning from others. 2. Download, with a prompt to choose a competition and download training data. 3. Build, with a prompt to build a model using preferred methods. 4. Submit, with a prompt to upload predictions and see the score on the leaderboard.

Active Competitions		Active Competitions	
All Competitions		<b>Springleaf Marketing Response</b> Determine whether to send a direct mail piece to a customer	15 days 1011 teams 1071 scripts \$100,000
		<b>Western Australia Rental Prices</b> Predict rental prices for properties across Western Australia	57 days 44 teams \$100,000

# Задачі статистичного аналізу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення








Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

## All Competitions

	<b>Springleaf Marketing Response</b> Determine whether to send a direct mail piece to a customer	15 days 1911 teams 1071 scripts \$100,000
	<b>Western Australia Rental Prices</b>  Predict rental prices for properties across Western Australia	57 days 44 teams \$100,000
	<b>Rossmann Store Sales</b> Forecast sales using store, promotion, and competitor data	2 months 362 teams 109 scripts \$35,000
	<b>Flavours of Physics: Finding <math>\tau \rightarrow \mu\mu</math></b> Identify a rare decay phenomenon	8.2 days 645 teams 602 scripts \$15,000
	<b>Truly Native?</b> Predict which web pages served by StumbleUpon are sponsored	10 days 205 teams \$10,000
	<b>Right Whale Recognition</b>	3 months 111 teams

# Задачі статистичного аналізу

Імовірнісні  
основи  
обробки  
даних

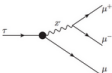
Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



A Feynman diagram illustrating the decay of a tau lepton ( $\tau$ ) into a muon ( $\mu$ ) and a muon-antimuon pair ( $\mu^+\mu^-$ ). The tau lepton enters from the left, emits a virtual photon ( $\gamma^*$ ), and then decays into a muon. The virtual photon subsequently decays into a muon-antimuon pair.

www.kaggle.com/c/flavours-of-physics

kaggle

Host Competitions Scripts Jobs Community

Sign up Login

\$15,000 • 645 teams

## Flavours of Physics: Finding $\tau \rightarrow \mu\mu\mu$

Mon 20 Jul 2015

Merger and 1st Submission Deadline

Mon 12 Oct 2015 (8.2 days to go)

### Dashboard

- Home
- Data
- Make a submission

### Information

- Description
- Evaluation
- Rules
- Prizes
- About the Sponsors
- Agreement test

### Competition Details » Get the Data » Make a submission

## Identify a rare decay phenomenon

Like last year's [Higgs Boson Machine Learning Challenge](#), this competition deals with the physics at the [Large Hadron Collider \(LHC\)](#). However, the subject of last year's challenge, the Higgs Boson, was already known to exist. The aim of this year's challenge is to find a phenomenon that is not already known to exist – charged lepton flavour violation – thereby helping to establish "new physics".

# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Номер проблемного места	Описание проблемы	Действие	Участок	Знание	Расположение	Дата открытия	Дата закрытия	Владелец	Категория
243	Голубой ящик для инструментов перегружен, и инструменты расположены хаотично	Навести порядок в ящике и избавиться от ненужных предметов	CDB*	4	Цех	24.02.00	18.03.00	J.D.	Расчистка/расстановка
845	Цель на подъемной двери заржавела	Заменить цель	CDB	4	Помещение для хранения окислителей	25.07.00	5.09.00	G.A.	Поддержание порядка
252	Фильтровальные установки расставлены беспорядочно	Заново расставить их в соответствии с размером, формой и предназначением	CDB	4	Цех	24.02.00	18.03.00	J.D.	Поддержание порядка/расстановка
1952	Оборудование расположено слишком близко к аппарату для промывки глаз	Переместить оборудование на расстояние трех футов от аппарата	CDB	4	Комната для производства подложек	12.07.01	12.07.01	Команда C	Безопасность
843	Вода разбрызгивается на пол около оборудования	Установить защитный экран для перенаправления потока воды	CDB	4	Комната для производства подложек	28.07.00	14.08.00	S.S.	Непрерывное улучшение



# Подання даних

Імовірнісні  
основи  
обробки  
даних

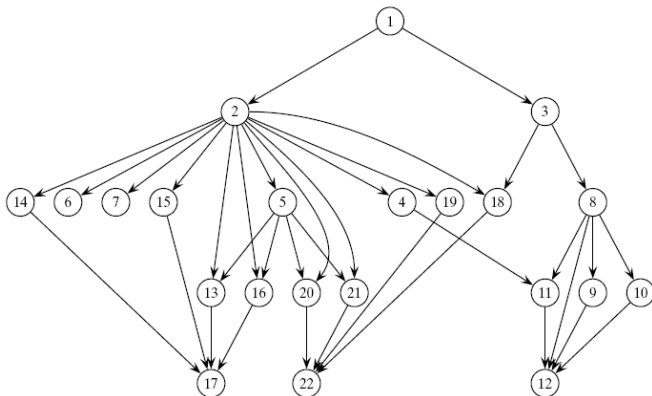
Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Table 1.1. Extract from the Iris dataset

	Sepal length	Sepal width	Petal length	Petal width	Class
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
$x_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$x_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$x_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$x_4$	4.6	3.2	1.4	0.2	Iris-setosa
$x_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$x_6$	4.7	3.2	1.3	0.2	Iris-setosa
$x_7$	6.5	3.0	5.8	2.2	Iris-virginica
$x_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$x_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

# Подання даних

Імовірнісні  
основи  
обробки  
даних

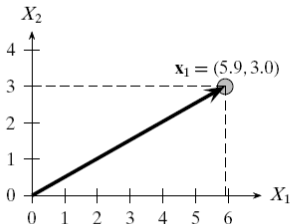
Основні  
визначення

Задачі ста-  
тистичного  
аналізу

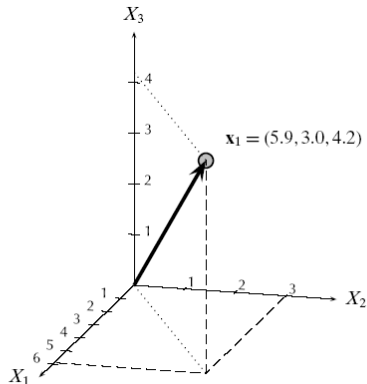
Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



(a)



(b)

# Подання даних

Імовірнісні  
основи  
обробки  
даних

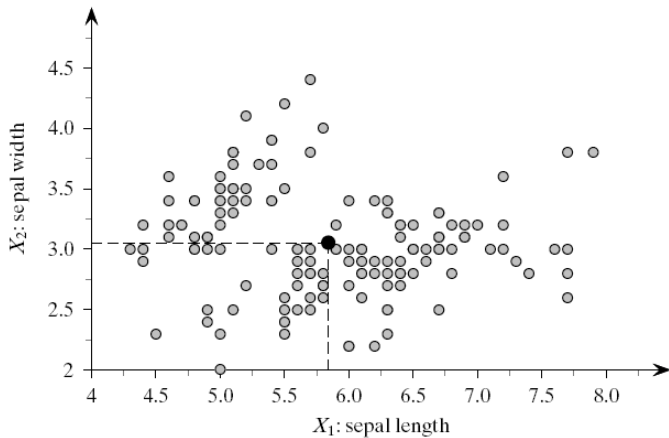
Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	4.9	7.0	5.5	6.3	6.8	6.1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	5.0	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.7	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	6.9	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1

# Подання даних

Імовірнісні  
основи  
обробки  
даних

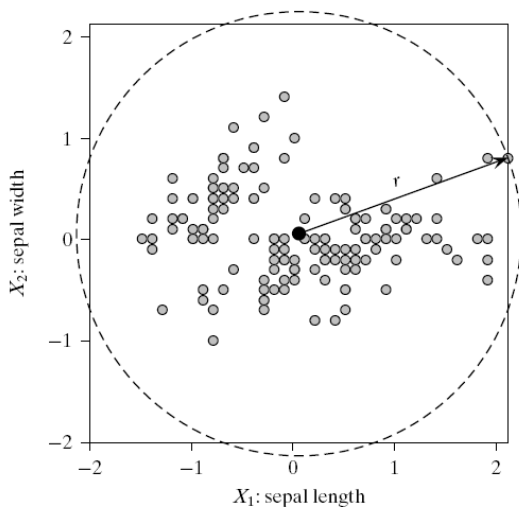
Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



# Подання даних

Імовірнісні  
основи  
обробки  
даних

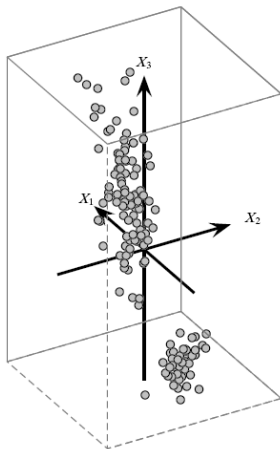
Основні  
визначення

Задачі ста-  
тистичного  
аналізу

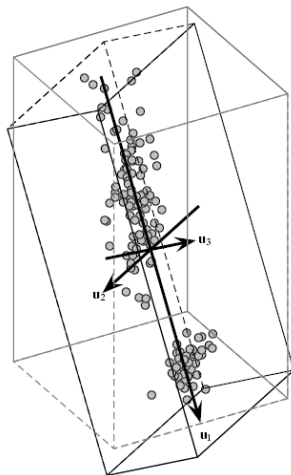
Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



(a) Original Basis



(b) Optimal Basis



# Подання даних

Імовірнісні  
основи  
обробки  
даних

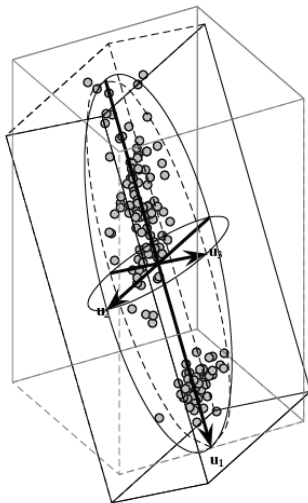
Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди



(a) Elliptic contours in standard basis

# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

<b>D</b>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

(a) Binary database

<i>t</i>	<b>i(t)</b>
1	<i>ABDE</i>
2	<i>BCE</i>
3	<i>ABDE</i>
4	<i>ABCE</i>
5	<i>ABCDE</i>
6	<i>BCD</i>

(b) Transaction database

<i>x</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<b>t(x)</b>	1	1	2	1	1
	3	2	4	3	2
	4	3	5	5	3
	5	4	6	6	4
		5			5
		6			

(c) Vertical database

# Подання даних

## Граф телефонних розмов



Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

## Подання даних

Імовірнісні  
основи  
обробки  
даних

## Подання даних

Номер проблемного места	Описание проблемы	Действие	Участок	Знание	Расположение	Дата открытия	Дата закрытия	Владелец	Категория
243	Голубой ящик для инструментов перетружен, и инструменты расположены хаотично	Навести порядок в ящике и избавиться от ненужных предметов	CDB*	4	Цех	24.02.00	18.03.00	J.D.	Расчистка/ расстановка
845	Цепь на подъемной двери заржавела	Заменить цепь	CDB	4	Помещение для хранения окислителей	25.07.00	5.09.00	G.A.	Поддержание порядка
252	Фильтровальные установки расставлены беспорядочно	Заново расставить их в соответствии с размером, формой и предназначением	CDB	4	Цех	24.02.00	18.03.00	J.D.	Поддержание порядка/ расстановка
1952	Оборудование расположено слишком близко к аппарату для промывки глаз	Переместить оборудование на расстояние трех футов от аппарата	CDB	4	Комната для производства подложек	12.07.01	12.07.01	Команда С	Безопасность
843	Вода разбрызгивается на пол около оборудования	Установить защитный экран для перенаправления потока воды	CDB	4	Комната для производства подложек	28.07.00	14.08.00	S.S.	Непрерывное улучшение

# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

	Наличие человека Н+	Отсутствие человека Н-	Сумма
Положительный тест Т+	14	4	18
Отрицательный тест Т-	2	10	12
Сумма	16	14	30

# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

	A	B	C	D
1	Test	Human		
2	1	1		
3	1	0		
4	0	0		
5	1	1		
6	1	1		

# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Ранжированный вариационный ряд распределения		Дискретный вариационный ряд распределения	
Количество полученных книг	Число студентов, получивших такое количество книг	Количество полученных книг	Доля студентов в общей совокупности
2	7	2	$7/40 = 0,175$
3	9	3	$9/40 = 0,225$
4	9	4	$9/40 = 0,225$
5	5	5	$5/40 = 0,125$
6	6	6	$6/40 = 0,150$
7	3	7	$3/40 = 0,075$
10	1	10	$1/40 = 0,025$
Итого:	40	Итого:	1

# Подання даних

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Среднедушевой доход в месяц, тыс. р.	Население	
	млн чел.	в % к итогу
До 400,0	29,0	17,4
400,1—600,0	29,1	15,6
600,1—800,0	24,2	13,5
800,1—1000,0	18,0	10,9
1000,1—1200,0	12,9	8,5
1200,1—1600,0	15,7	11,8
1600,1—2000,0	8,1	7,3
Свыше 2000,0	10,5	15,0
Итого	147,5	100,0



# Подання даних

Імовірнісні  
основи  
обробки  
даних

Table 19.1. Discretized `sepal length` attribute: class frequencies

Bins	$v$ : values	Class frequencies ( $n_{vi}$ )	
		$c_1$ :iris-setosa	$c_2$ :other
[4.3, 5.2]	Very Short ( $a_1$ )	39	6
(5.2, 6.1]	Short ( $a_2$ )	11	39
(6.1, 7.0]	Long ( $a_3$ )	0	43
(7.0, 7.9]	Very Long ( $a_4$ )	0	12

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

# Оцінки параметрів розподілу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Нехай  $X_n = \{x_1, x_2, \dots, x_n\}$  — вибірка об'ємом  $n$  з генеральної сукупності значень випадкової величини  $X$  з середнім значенням  $\bar{x}$  (математичним очікуванням  $M[X]$ ), дисперсією  $\sigma^2$  ( $D[X]$ ) і середньоквадратическим відхиленням  $\sigma = \sqrt{\sigma^2} = \sqrt{D[X]}$ .

# Оцінки параметрів розподілу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Нехай  $X_n = \{x_1, x_2, \dots, x_n\}$  — вибірка об'ємом  $n$  з генеральної сукупності значень випадкової величини  $X$  з середнім значенням  $\bar{x}$  (математичним очікуванням  $M[X]$ ), дисперсією  $\sigma^2$  ( $D[X]$ ) і середньоквадратическим відхиленням  $\sigma = \sqrt{\sigma^2} = \sqrt{D[X]}$ .
- **Вибірковим середнім** вибірки називається середнє арифметичне

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

# Оцінки параметрів розподілу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- **Вибірковою дисперсією** називається

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

# Оцінки параметрів розподілу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- **Вибірковою дисперсією** називається

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

- **Модифікованою вибірковою дисперсією** називається

$$\sigma_m^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

# Оцінки параметрів розподілу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Всі ці вибіркові величини залежать від вибірки і самі є випадковими величинами, їх значення лише наближено дорівнюють відповідним числовим характеристикам генеральної сукупності.

# Оцінки параметрів розподілу

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Всі ці вибірккові величини залежать від вибірки і самі є випадковими величинами, їх значення лише наближено дорівнюють відповідним числовим характеристикам генеральної сукупності.
- Статистикою називається будь-яка функція, що залежить від вибірки і сама є випадковою величиною. Таким чином, вибірккове середнє  $\bar{X}$ , вибірккова дисперсія  $\sigma^2$  ( $D[X]$ ) і модифікована вибірккова дисперсія  $\sigma_m^2$  — це статистики.

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- **Статистичним рядом** називається сукупність пар  $i \implies x_i$ , отриманих в результаті експерименту. Зазвичай статистичні ряди оформляються у вигляді таблиці (таблиця 2), в першому стовпці якої стоїть номер дослідів ( $i$ ), а в другому — спостережуване значення випадкової величини  $x_i$ , яке називається **варіантою**.



# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- **Статистичним рядом** називається сукупність пар  $i \Rightarrow x_i$ , отриманих в результаті експерименту. Зазвичай статистичні ряди оформляються у вигляді таблиці (таблиця 2), в першому стовпці якої стоїть номер досліду ( $i$ ), а в другому — спостережуване значення випадкової величини  $x_i$ , яке називається **варіантою**.

Индекс $i$	Варианта $x_i$
1	$x_1$
2	$x_2$
...	...
$n$	$x_n$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

**Розмахом вибірки** називають різницю між найбільшою і найменшою варіантами вибірки:

$$R = x_{max} - x_{min}.$$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Якщо одна і та ж варіанта зустрічається у вибірці кілька разів, то статистичний ряд зручніше записувати у вигляді таблиці

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Таблиця 2

Індекс $i$	Варіанта $x_i$	Частота $n_i$	Относит. частота $\bar{n}_i$
1	$x_1$	$n_1$	$\bar{n}_1$
2	$x_2$	$n_2$	$\bar{n}_2$
...	...	...	...
$k$	$x_k$	$n_k$	$\bar{n}_k$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Частотою  $n_i (i = \overline{1, k})$  варіанти  $x_i$  називається число повторень варіанти  $x_i$  у вибірці, причому

$$\sum_{i=1}^k n_i = n.$$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Частотою  $n_i (i = \overline{1, k})$  варіанти  $x_i$  називається число повторень варіанти  $x_i$  у вибірці, причому

$$\sum_{i=1}^k n_i = n.$$

- Відносною частотою або вагою  $\bar{n}_i (i = \overline{1, k})$  варіанти  $x_i$  називається відношення частоти варіанти до об'єму вибірки  $n$ , тобто

$$\bar{n}_i = \frac{n_i}{n}$$

$$\sum_{i=1}^k \bar{n}_i = 1.$$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

При великій кількості спостережень простий статистичний ряд перестає бути зручною формою запису статистичних даних. Для додання йому більшої компактності і наочності статистичний матеріал піддають додатковій обробці — будують варіаційні ряди або груповані варіаційні ряди.

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Варіаційним рядом називається упорядкована сукупність варіант  $x_i (i = \overline{1, k})$  з відповідними їм частотами  $n_i$  або відносними частотами  $\bar{n}_i$ .

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Варіаційним рядом називається упорядкована сукупність варіант  $x_i (i = \overline{1, k})$  з відповідними їм частотами  $n_i$  або відносними частотами  $\bar{n}_i$ .
- Для побудови групованого варіаційного ряду інтервал зміни спостережених значень випадкової величини  $[x_{min}; x_{max}]$  розбивають на  $N$  інтервалів, що не пересікаються (їх називають частковими інтервалами або розрядами).



# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Число інтервалів залежить від об'єму вибірки і визначається за формулою Стерджеса



$$N = 1 + 3.32 \log n$$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Число інтервалів залежить від об'єму вибірки і визначається за формулою Стерджеса



$$N = 1 + 3.32 \log n$$



$$N = 1 + 1.44 \ln n$$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Число інтервалів залежить від об'єму вибірки і визначається за формулою Стерджеса



$$N = 1 + 3.32 \log n$$



$$N = 1 + 1.44 \ln n$$



$$N \geq [1 + 3.32 \log n] + 1$$

квадратні дужки позначають цілу частину числа.

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Розбиття на мале число інтервалів може призвести до невірних статистичними висновків. Відповідно до цієї формули, необхідно брати не менше 8 інтервалів на 100 спостережень.

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Розбиття на мале число інтервалів може призвести до невірних статистичними висновків. Відповідно до цієї формули, необхідно брати не менше 8 інтервалів на 100 спостережень.
- Інтервали повинні бути однакової довжини

$$\Delta = \frac{R}{N} = \frac{x_{max} - x_{min}}{N}$$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Частотою  $n_i (i = \overline{1, N})$  інтервалу  $(u_i; u_{i+1}]$  називається число варіант  $x_i$ , що потрапили в цей інтервал, причому

$$\sum_{i=1}^N n_i = n.$$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- Частотою  $n_i (i = \overline{1, N})$  інтервалу  $(u_i; u_{i+1}]$  називається число варіант  $x_i$ , що потрапили в цей інтервал, причому

$$\sum_{i=1}^N n_i = n.$$

- При групуванні спостережених значень за розрядами виникає питання про те, до якого інтервалу віднести значення, що знаходиться на кордоні двох розрядів. В цих випадках вважають дане значення належить до лівого інтервалу.

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- **Частотою**  $n_i (i = \overline{1, N})$  інтервалу  $(u_i; u_{i+1}]$  називається число варіант  $x_i$ , що потрапили в цей інтервал, причому

$$\sum_{i=1}^N n_i = n.$$

- При групуванні спостережених значень за розрядами виникає питання про те, до якого інтервалу віднести значення, що знаходиться на кордоні двох розрядів. В цих випадках вважають дане значення належить до лівого інтервалу.
- **Відносною частотою** або **вагою**  $\bar{n}_i (i = \overline{1, N})$  інтервалу  $(u_i; u_{i+1}]$  називається відношення частоти інтервалу до об'єму вибірки  $n$ , тобто

$$\bar{n}_i = \frac{n_i}{n}$$



Імовірнісні  
основи  
обробки  
даних

Г

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- **Накопиченою відносною частотою**  $w_i (i = \overline{1, N})$  інтервалу  $(u_i; u_{i+1}]$  називається сума відносних частот перших  $i$  інтервалів, тобто

$$w_i = \sum_{j=1}^i \bar{n}_j.$$

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

- **Накопиченою відносною частотою**  $w_i (i = \overline{1, N})$  інтервалу  $(u_i; u_{i+1}]$  називається сума відносних частот перших  $i$  інтервалів, тобто

$$w_i = \sum_{j=1}^i \bar{n}_j.$$

- **Групованим варіаційним рядом** називається впорядкована сукупність інтервалів з відповідними їм частотами  $n_i$ , відносними частотами  $\bar{n}_i$  і накопиченими відносними частотами  $w_i$ .

# Статистичні ряди

Імовірнісні  
основи  
обробки  
даних

Основні  
визначення

Задачі ста-  
тистичного  
аналізу

Подання  
даних

Оцінки  
параметрів  
розподілу

Статистичні  
ряди

Індекс $i$	Інтервал $(u_i; u_{i+1}]$	Частота $n_i$	Относит. частота $\bar{n}_i$	Накопл. относит. частота $w_i$
1	$[u_1; u_2]$	$n_1$	$\bar{n}_1$	$w_1 = \bar{n}_1$
2	$(u_2; u_3]$	$n_2$	$\bar{n}_2$	$w_2 = \bar{n}_1 + \bar{n}_2$
...	...	...	...	...
$N$	$(u_N; u_{N+1}]$	$n_N$	$\bar{n}_N$	$w_N = 1$
$\sum_{i=1}^N$		$n$	1	