**ПОЛТАВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ**
**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ЕКОНОМІКИ, УПРАВЛІННЯ, ПРААВ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**
**КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА ТЕХНОЛОГІЙ**

# Пояснювальна записка

до кваліфікаційної роботи на здобуття ступеня вищої освіти Бакалавр

на тему: «Розроблення моделі потоку задач в ГРІД та її програмної реалізації»

Виконав: здобувач вищої освіти
за освітньо-професійною програмою
Інформаційні управляючі системи
спеціальності 126 Інформаційні
системи та технології
освітнього ступеня Бакалавр
групи 126ІСТ_бд_41
Волошин Є.В.
Керівник: Одарущенко О.М.
Рецензент: Брикун О.М.

**Полтава – 2024 року**

# ВСТУП

*Актуальність теми.* Область розподілених обчислювальних систем в даний час характеризується швидкими темпами зміни ідеологій і підходів. За коротку історію існування систем такого типу з'явилося безліч різноманітних парадигм реалізації розподілених обчислень, що набрали великий увагу і загальне визнання, але практично зниклих згодом під тиском більш нових і модних підходів. Однак коли технологія зникає із уваги, дуже часто вона з'являється знову під новим ім'ям. У результаті відбувається безперервне перемішування базових концепцій з новітніми підходами до розробки.

У середині 1990-х існувало два основних підходи до розробки розподілених обчислювальних систем. З одного боку, концепція Веб представляла собою орієнтоване на людину розподілений інформаційний простір. З іншого боку, технології розподілених об'єктів, такі як CORBA і DCOM були в першу чергу орієнтовані на створення розподілених середовищ, які б емулюватися процес розробки і використання локальних додатків, забезпечуючи переваги доступу до мережевих ресурсів. Але, незважаючи на початкову ідею Веб як простору, який дозволяв би багатьом людям обмінюватися інформацією, не публікуючи нічого в свою чергу. Між іншим системи розподілених об'єктів росли з погляду наданих їм можливостей, але ставали все більш важкими в плані розробки і використання.

Відразу після початку нового тисячоліття стався вибух розвитку нових методів і проміжного програмного забезпечення для розподілених обчислювальних систем, включаючи технології однорангових мереж (peer-to-peer або P2P) і грід-технології. Застосування P2P дозволило безлічі користувачам, які раніше були простими споживачами інформації, взяти участь у наданні контенту. З іншого боку, застосування технології ГРІД дозволило інтегрувати великі комплекси обробки і зберігання даних, забезпечуючи їх доступність для різних урядових і наукових користувачів.

ГРІД – це форма розподілених обчислень, в якій «віртуальний суперкомп'ютер» представлений у виді кластера, з'єднаних за допомогою мережі, комп'ютерів, що працюють разом для виконання наукових, математичних задач, що потребують значних обчислювальних ресурсів. За допомогою ГРІД виконуються деякі трудомісткі завдання, пов'язані з економічною прогнозуванням, сейсмоаналізом, розробкою та вивченням властивостей нових ліків.

Одним із механізмів для збільшення відмовостійкості у ГРІД є розуміння характеру потоку задач і його розподілу між ресурсами. Для побудови моделі потоку задач у даній роботі використовуються методи математичної статистики кількісного і графічного представлення даних у ГРІД, адекватність отриманих моделей перевіряється критерієм згоди Колмогорова.

*Мета роботи* – розглянути архітектуру ГРІД, дослідити методи і технології аналізу і моделювання потоку задач в ГРІД, розробити програмну реалізацію для статистичної обробки даних їх кількісного і графічного представлення, оцінки адекватності отриманої моделі робочого навантаження задач у ГРІД.

*Об'єкт дослідження* – процеси моделювання потоку задач в ГРІД.

*Предмет дослідження* – методи та засоби моделювання потоку задач в ГРІД.

*Методи дослідження.* Проведені в роботі дослідження ґрунтуються на методах теорії ймовірностей, системного, марківського аналізу та математичної статистики, імітаційного моделювання для побудови імітаційної моделі функціонування ВКС в умовах зміни параметрів потоків відмов та відновлення ПЗ.

*Інформаційна база* кваліфікаційної роботи сформована з наукових фахових статей, наукових монографій, аналітичних звітів державних та міжнародних експертних, виконаних науково-дослідних та дисертаційних робот за заданою тематикою.

*Практична значущість:* розроблено програмне забезпечення моделювання потоку задач в ГРІД.

Робота складається зі вступу, трьох розділів, висновків та списку літератури. Обсяг роботи становить 67 сторінок з додатками, 7 таблиць, 23 рисунків, 2-х додатків.

# РОЗДІЛ 1
# АНАЛІЗ ТА ОБРАННЯ ТЕХНОЛОГІЙ ГРІД

## 1.1 Основні поняття ГРІД

Термін «ГРІД» став використовуватися з середини 90-х років для позначення деякої інфраструктури розподіленого комп'ютингу, пропонованої для обслуговування передових наукових та інженерних проектів. Основною метою ГРІД є інтеграція великої кількості розподілених ресурсів, які можуть взаємодіяти один з одним для досягнення певної мети. Все більша кількість організацій, дослідницьких груп, учених, залучаються до ГРІД інфраструктури. У міру свого розвитку, технологія отримала значне поширення і з області фундаментальних наук перейшла у сферу бізнес рішень і послуг.

Однак з переходом на нову платформу з'явився ряд проблем, пов'язаних з розподілом навантаження в рамках гетерогенного середовища. Із збільшенням кількості учасників і ресурсів ГРІД, розподілених в різних просторових (географічне положення) і тимчасових рамках (часові пояси), виникає проблема рівномірного (необхідного) рівня навантаження розподілених ресурсів, зниження сумарного часу виконання задач і кількості відмов, підвищення продуктивності, як окремих вузлів , так і системи в цілому.

ГРІД (англ. ГРІД - решітка, мережа) - узгоджене, відкрите й стандартизоване комп'ютерне середовище, яке забезпечує гнучкий, безпечний, скоординований розподіл обчислювальних ресурсів і ресурсів зберігання інформації, які є частиною цього середовища, в рамках однієї ВО. (Я. Фостер, К. Кессельман).

ГРІД (англ. ГРІД - решітка, мережа) - система, пов'язана з функціями інтеграції, віртуалізації і управління службами та ресурсами в розподіленому, гетерогенному середовищі, яка підтримує сукупність користувачів і ресурсів (ВО) на сукупності традиційних адміністративних і організаційних доменів (фактичних організацій ).

ГРІД обчислення - це форма розподілених обчислень, в якій «супер і віртуальний комп'ютер» представлено у вигляді кластера, з'єднаних за допомогою мережі, слабозв’язаних комп'ютерів, що працюють разом для обробки величезної кількості задач (операцій, робіт).

Віртуальна організація (англ. Virtual Organization) - являється динамічною спільнотою людей та/чи установ, які спільно використовують обчислювальні ресурси відповідно до узгоджених між ними правилами. Ці правила регулюють доступ до всіх типів засобів, включаючи комп'ютери, програмне забезпечення та дані.

## 1.2 Архітектура ГРІД

Завдяки реалізації ГРІД на основі існуючих протоколів і стандартів, досягається її високий рівень інтероперабельності та масштабованості, що дозволяє розширювати систему у двох напрямках: вертикальному і горизонтальному (на рівні додатку).

Вертикальна архітектура, представлена на рисунку 1.1, дозволяє встановлювати ГРІД системи «поверх» існуючих програмних і апаратних рішень, що значно прискорює процес їх інтеграції. У свою чергу, розширення горизонтальної архітектури обумовлюється появою нових ГРІД служб і послуг.

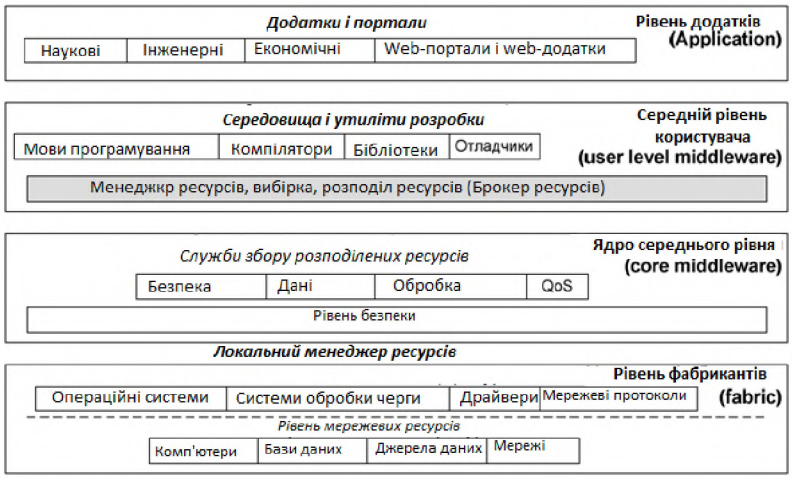



Рисунок 1.1 – Вертикальна архітектура ГРІД

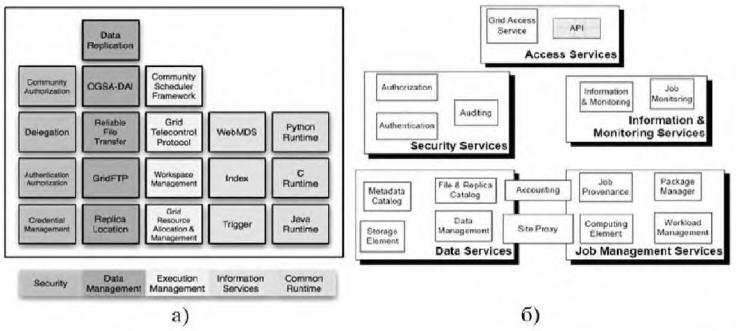
На сьогоднішній день стандарт OGSA (Open GRID Services Architecture) визначає такі основні групи служб, що функціонують на рівні додатків:

- сервіси управління виконанням завдань (Execution Management services);
- сервіси управління даними (Data services);
- сервіси управління ресурсами (Resource Management services);
- сервіси безпеки (Security services);
- інформаційні сервіси (Information services);
- сервіси самоуправління (Self-management services).

Найчастіше остання група служб (Self-management), інтегрується в інші п'ять і в явному вигляді стандарт не вимагає її окремої реалізації в ГРІД.

Існує досить велика кількість проектів і рішень в області ГРІД, що реалізують методи та інструментальні засоби підтримки функціонування ГРІД-систем. Розглянемо архітектуру ГРІД на прикладі двох програмних продуктів - Globus, який став стандартом де-факто і визнаний багатьма світовими виробниками комп'ютерної індустрії IBM, SGI, Sun Microsystems, Fujitsu, Hitachi, NEC, Veridian, Entropia, Platform Computing Inc , Microsoft, Compaq і gLite, що стартував у рамках Європейського проекту EGEE (Enabling GRIDs for E-Science).

На рисунку 1.2 наведені розглянуті раніше групи служб, що формують ядро ГРІД, які в свою чергу складаються з набору окремих сервісів (горизонтальна архітектура), що реалізують функціональність своєї групи (на прикладі Globus і gLite).




а) б)

Рисунок 1.2 – Горизонтальна архітектура ГРІД:

а) Globus; б) gLite

У загальному випадку вертикальна і горизонтальна архітектура можуть бути представлені наступним чином (див. рис. 1.3).

Таким чином, будь-яка обчислювальна одиниця або ресурс (кластер, окремий комп'ютер, сервер баз даних і т.д.), що підключається до ГРІД, повинен:

- підтримувати вертикальну і горизонтальну архітектуру ГРІД;
- належати одній або більше групам служб. Стандарт не визначає обмеження на кількість служб, встановлених на одному комп'ютері, єдине обмеження - це допустиме зниження продуктивності;
- належати одній або більше ВО, в рамках яких він буде функціонувати.

ВО - це динамічно формована сукупність окремих користувачів, груп і установ, які визначили умови і правила поділу ресурсів. Концепція ВО є ключовою для ГРІД-комп'ютингу. Всі ВО беруть участь у виробленні угод за характеристиками і спірних питань, що включає в себе загальні інтереси і потреби, які можуть змінюватись за обсягом, області дії, часу виконання, соціологічним параметрами і структурі. Учасники будь-якої ВО домовляються про спільний поділ ресурсів, заснованому на правилах і умовах, визначених цією ВО і потім отримують доступ до ресурсів у створеному пулі ВО.

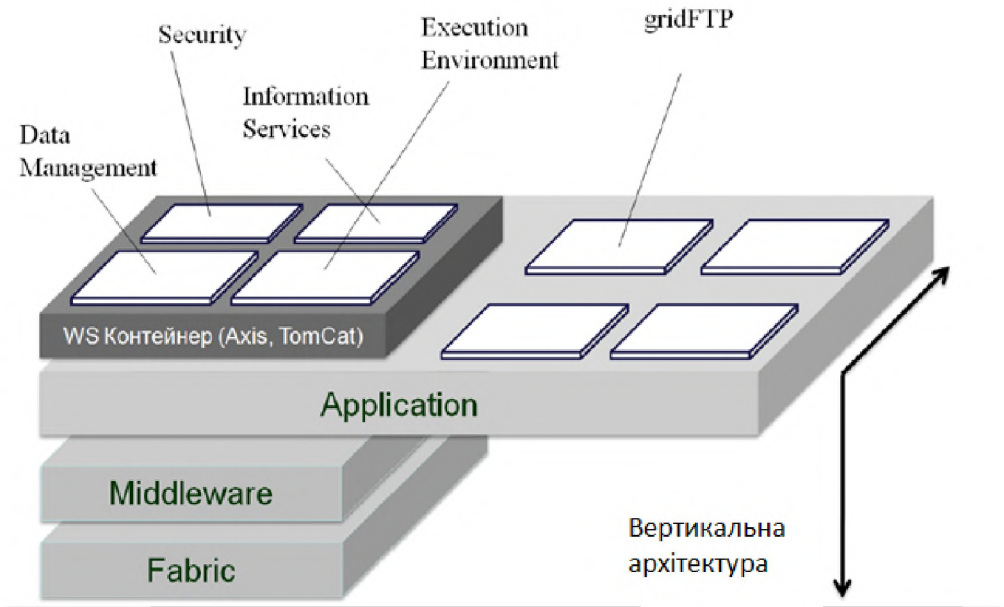



Рисунок 1.3 – Архітектура ГРІД системи

Таким чином, для доступу до розподіленої системи незалежно від ролі учасника (користувач або обчислювальний вузол, «виробник/споживач») необхідною умовою є приналежність до тієї чи іншої ВО.

На підставі аналізу структури ГРІД доцільно виділити три основні позиції, які необхідно враховувати при побудові її моделі:

1. ВО не позначають володіння ресурсами, визначеними у рамках їх функціонування, вони лише встановлюють правила взаємодії. Тобто, в рамках ВО ресурси можуть належати різним адміністративним установам.

2. ВО володіє повною інформацією про кількість ресурсів і користувачів, зареєстрованих в ній, однак не гарантує їх доступність і тимчасову актуальність.

3. один і той же ГРІД-ресурс може використовуватися декількома ВО.

## 1.3 Розподіл ресурсів і потоків задач в ГРІД

Під ресурсом у ГРІД мають на увазі сукупність апаратного (процесори, пам'ять, мережеве обладнання тощо) та прикладного (системного) ПЗ (програми, бібліотеки функцій і т.д.), об'єднаного в єдину структуру і наданого для вирішення різного типу задач .

Аналіз ресурсів ГРІД дозволив позначити чотири основних групи ресурсів:

- обчислювальні ресурси – окремі комп’ютери, кластери;

- ресурси збереження даних – диски і дискові масиви, стрічки, системи масового збереження даних;

- ПЗ (програми, бібліотеки, ін. )

На основі аналізу архітектури ГРІД в рамках існуючих проектів, і стандартів ГРІД можна виділити наступні типи ресурсів (рисунок 7):

- сайт (Site), що складається з елементів:

1) комп'ютер кінцевого користувача (User Interface, UI). Це комп'ютер, на якому встановлено програмні засоби для користувацького інтерфейсу і який

дозволяє кінцевому користувачеві взаємодіяти з ГРІД-середовищем (зокрема, запускати задачі і отримувати результати);

2) обчислювальний елемент (Computing Element, CE). Обчислювальний елемент являє собою ГРІД-інтерфейс для локальної системи управління пакетної обробки (СУПО) (або локального менеджера ресурсів (ЛМР)), зазвичай виступає в ролі посередника між локальним обчислювальним кластером і ГРІД;

3) робочі вузли (Woker Nodes, WN). З точки зору ГРІД-середовища робочі вузли знаходяться за обчислювальним елементом (РЕ) і керуються локальною СУПО. Деталі процесу розподілу та обчислення виявляються прихованими для кінцевого користувача, але саме ці вузли виконують фактичні обчислення і, значить, на них повинно бути встановлено програмне забезпечення для виконання завдань кінцевих користувачів;

4) накопичувач даних (Storage Element, SE). Цей вузол забезпечує однаковий доступ до будь-яких накопичувачам даних. У загальному випадку, накопичувач може керувати дисковими масивами, масовою пам'яттю і т.п. Цей елемент приховує деталі конкретної накопичувальної системи і забезпечує користувачам однаковий доступ до даних.

- окремий вузол, який реалізовує:

1) брокер ресурсів (Resource Broker, RB). Цей вузол приймає завдання від користувача (через інтерфейс користувача), узгоджує вимоги до ресурсів, що містяться в описі завдання, з наявними в наявності вільними ресурсами і направляє завдання на відповідний сайт;

2) каталог реплік (Replica Catalog, RC). Цей елемент підтримує базу даних про місця зберігання оригінальних файлів і всіх їх копій;

3) інформаційний сервіс (Information Service, IS). Вузол відповідає за отримання, обробку та надання інформації про стан ГРІД ресурсів.

Загальна структура взаємодії ресурсів в ГРІД представлена на рисунку 1.4. Кількість ресурсів і сервісів, які вони реалізують, залежить від інфраструктури і призначення ГРІД - визначається політикою ВО, безліччю сайтів системи в цілому, і в окремо взятій ВО (наприклад, проект LCG об'єднує 185 сайтів, 39 ВО,

31 RB). Велика кількість RB дозволяє збільшити кількість «точок входу» в ГРІД, однак вимагає більшої узгодженості в управлінні ресурсами, занадто мала кількість RB може стати «вузьким» місцем у системі і знизити продуктивність ГРІД в цілому.

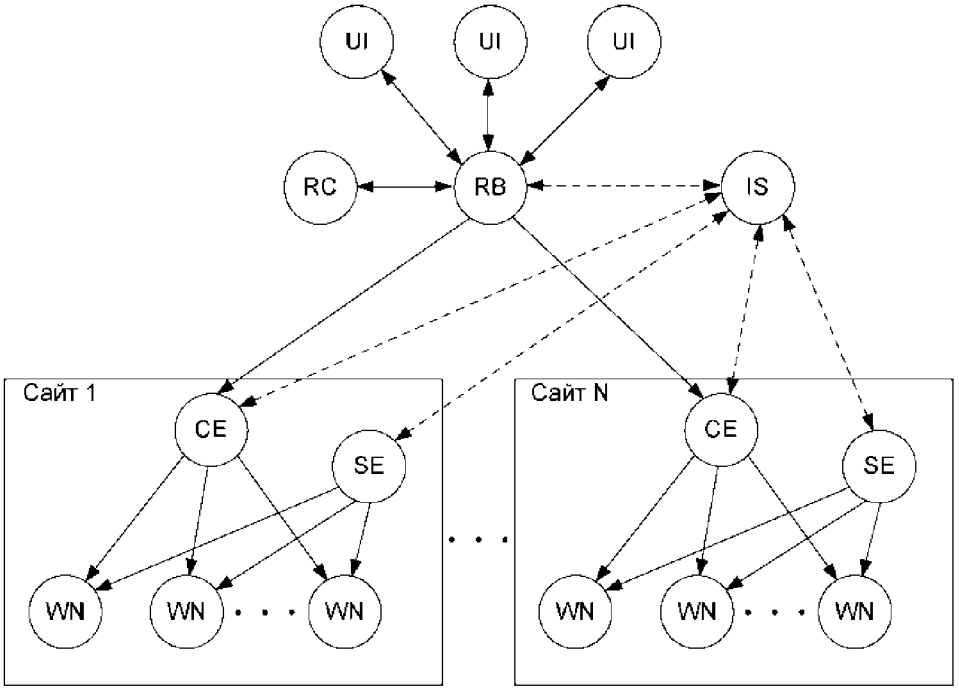



Рисунок 1.4 – Структурна схема ГРІД

Як правило, доступ до ресурсів визначається набором політик, прийнятих у ВО, і здійснюється за коштами механізму сервісів, об'єднаних у функціональні групи, розглянутих нами раніше.

Основне призначення ГРІД  виконання обчислень (обробка великого обсягу даних, моделювання, прогнозування і т.д.), тобто задачі, що вимагають великої кількості обчислювальних ресурсів і значних часових витрат.

OGSA визначає термін «задача» в рамках ГРІД, як задана користувачем одиниця роботи, яку необхідно виконати для досягнення поставленої мети (отримання результату). Задача, у свою чергу, є мінімальною одиницею, якою оперує ГРІД-система.

Аналіз типів задач у ГРІД показав, що при побудові моделі потоку задач доцільно враховувати дві основні складові: характер потоку задачі (структура) та вимоги, які пред'являються задачею до виконавчого середовищі.

Сукупність задач, із заздалегідь визначеною послідовністю виконуваних кроків і механізмами взаємодії між ними, спрямованих на вирішення загальної проблеми (отримання конкретного результату), формує потік задачі (job workflow) або ГРІД додаток (application) (найчастіше в ГРІД застосовують термінологію «задача, розбита на підзадачі »).

Виділяють наступні структури опису потоку завдання, які характеризують часову і функціональну залежність між підзадачами (послідовність виконання):

а) ациклічний орієнтований граф (directed acyclic graph, DAG):

1) послідовне виконання (squence);

2) паралельне виконання (parallelism);

3) виконання на основі умов (choice).

б) орієнтовані графи, які допускають цикли (ітерації).

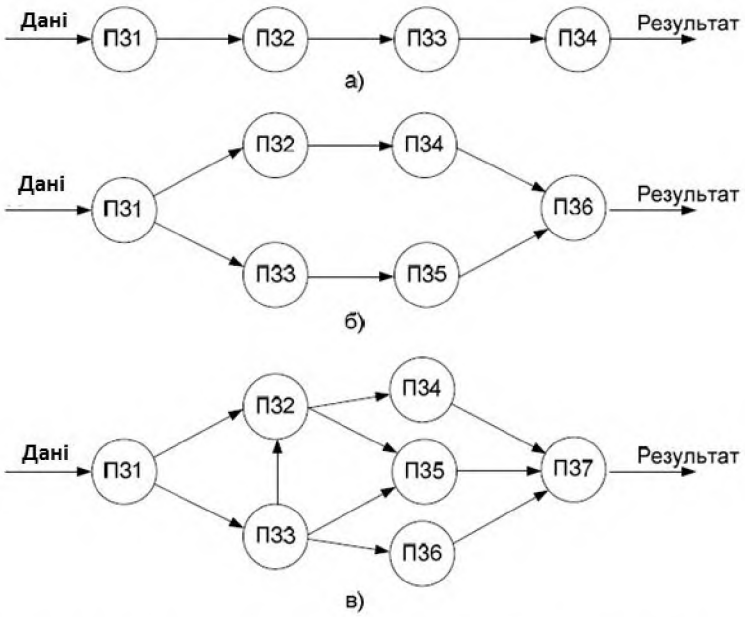



Рисунок 1. 5 –Типи задач в ГРІД: а) послідовний; б) паралельний; в) мережевий

Таким чином, можна виділити наступні типи потоків задачі ( рис. 1.5):

- послідовний (sequential) (рис. 1.5, а);

- паралельний (parallel) (рис. 1.5, б);

- змішаний або мережевий (network) (рис. 1.5, в).

Зазвичай задача, що використовує переваги ГРІД, розбивається на підзадачі, які, у свою чергу, виконуються незалежно одна від одної (на різних ресурсах), а результати їх обчислень передаються наступній по ланцюжку підзадачі. Задача вважається успішно вирішеною, якщо всі підзадачі були виконані успішно.

Другою складовою, при аналізі задачі в ГРІД, є вимоги, які пред'являються для ресурсів. Стандарт опису задачі в ГРІД JSDL (Job Sumiton Description Language) передбачає зазначення необхідних характеристик середовища виконання: архітектура процесора, тип файлової системи, тип операційної системи, обсяг необхідної оперативної пам'яті, джерело вхідних даних для обробки, необхідне програмне забезпечення і т. д.

Безліч задач різного типу (послідовних, паралельних, мережевих), що надходять в ГРІД систему в довільні проміжки часу, формують ПЗ або робоче навантаження (job workload).

Прийняття рішення про скоординований розподіл динамічних ресурсів і розподіл безлічі задач в умовах великої кількості ВО (а як результат і правил) формує основну проблему ГРІД обчислень.

Незважаючи на складну структуру ГРІД, система планування має досить просту архітектуру і являє собою програмний комплекс, що забезпечує збір інформації про стан і тип ресурсів, характер і тип задач; пошук і вибір необхідних ресурсів згідно деякої цільової функції і вимог, що пред'являються задачею; постановку задачі на виконання

Для ГРІД систем характерний наступний механізм взаємодії: планувальник отримує завдання від користувачів, вибирає необхідні ресурси для запуску задачі, використовуючи інформаційний сервіс ГРІД (наприклад, Monitor and Discovery System MDS), і в результаті направляє задачу на відповідний ресурс. Рішення про вибір того чи іншого ресурсу базується на вимогах самої задачі,

цільової функції або критерію (мінімізація часу розрахунку, надійність результату і т.д.) і стану системи (рис. 1.6).

Прикладом диспетчера ГРІД може бути Globus Grid Resource Allocation and Management. Локальний менеджер ресурсів, у свою чергу, часто являє собою систему управління пакетної обробки (наприклад, Portable Batch System, Condor-G, Load Sharing Facility, Maui Cluster Scheduler і т.д. ), яка об'єднує локальні ресурси в обчислювальний кластер і відповідає за розподіл задач, що надходять від локальних або ГРІД користувачів.

Таким чином, з точки зору архітектури ГРІД, механізм планування задач реалізований на двох рівнях: локальному і глобальному (рис. 1.7), де локальні задачі виконуються на рівні окремих обчислювальних кластерів і не поступають на другі системи, а глобальні задачі, можуть попадати на будь-який кластер, що входить у склад ГРІД.

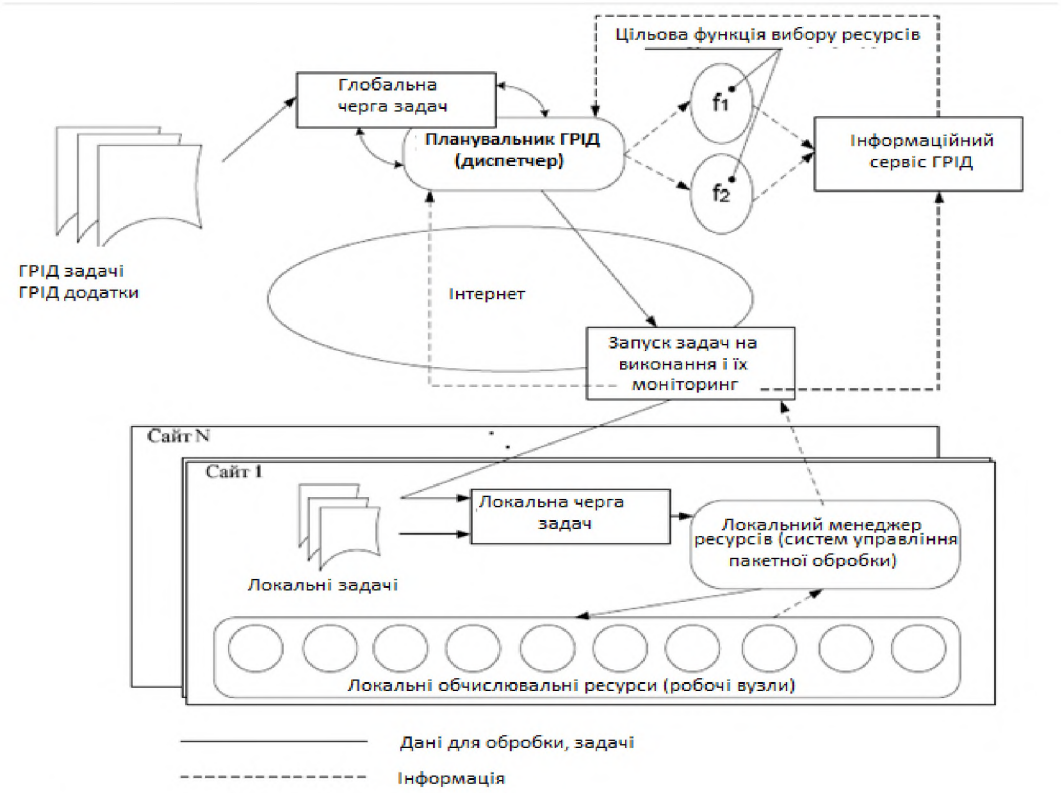



Рисунок 1.6 – Структурна схема планування задач і ресурсів

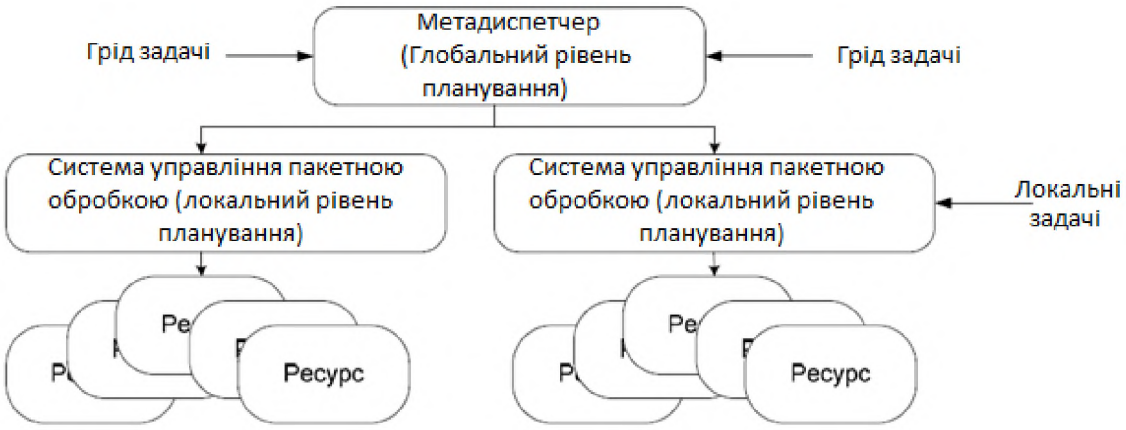



Рисунок 1.7 – Локальний і глобальний рівень координації задач і ресурсів

Відомо, що складність проблеми планування ресурсів відноситься до класу NPповних задач і тому точного вирішення їх немає. У зв'язку з цим для її вирішення використовуються класи наближених, евристичних алгоритмів і алгоритмів скорочення перебору.

## 1.4 Обґрунтування і вибір показників оцінки якості обслуговування у відмовостійких ГРІД системах

Аналіз показників якості обслуговування (Quality of Service, QoS) у ГРІД дозволяє зробити висновок про наявність трьох базових класифікаційних моделей:

1. модель метрик і політик (metrics and policies), де метрики грають роль кількісних показників, а політики визначають поведінкову роль компонентів системи;

2. метрики QoS на основі архітектурного підходу;

3. розширені метрики QoS базуються на моделі «метрик і політик» і доповнені за рахунок введення базових показників надійності. У розглянутій ієрархії (рис. 1.8) метрики якості обслуговування діляться на п'ять основних

класів: продуктивність, надійність (гарантоздатність), вартість, конфігурація і призначені для користувача (зумовлені користувачем) .

Більшість робіт, спрямованих на підвищення якості обслуговування в ГРІД системах, базуються на мінімізації або максимізації значення однієї або декількох метрик, наприклад, продуктивності ресурсів, доступності ресурсів, часу виконання задач, пропускної здатності каналу зв'язку і т.д.

У виду стохастичною природи ГРІДоточення, найбільш важливими є метрики продуктивності, зокрема, часові показники (час виконання задачі, час доступу до ресурсу, час очікування задачі в черзі), доступність ресурсів, завантаженість ресурсу (обсяг поступаючих задач за одиницю часу), інтенсивність відмов ресурсів.

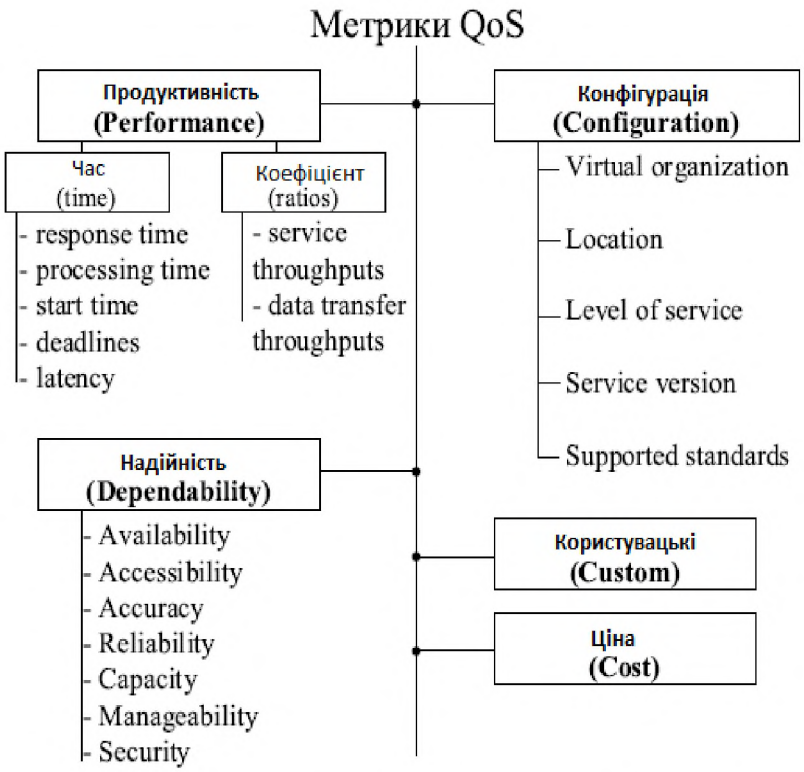



Рисунок 1.8 – Метрики QoS в ГРІД

Для підвищення відмовостійкості ГРІД можна використовувати різні показники, наведені на рисунку 1.8, зокрема:

- аналіз та прогнозування характеру потоку задач (workload characterization and prediction):

1) інтенсивності потоку задач(arrival rate);

2) пропускної здатності системи (throughput);

- прогнозування часу обслуговування задач (jobs runtime prediction);

- прогнозування завантаження системи (system utilization prediction);

- прогнозування появи відмови (fault forecasting) виконання задачі з причини недоступності ГРІДресурсів, тобто прогнозування відмови ресурсу (resource unavailability prediction).

В основу методології моделювання відмовостійких ГРІДсистем покладено принципи системного аналізу, теорії систем масового обслуговування, теорії ймовірності та математичної статистики.

## 1.5 Метологія аналізу і побудови моделі потоку задач в ГРІД

На основі аналізу архітектури ГРІДсистеми (рис. 1.3) і структурної схеми планувальника задач (рис. 1.6) доцільно виділити наступну послідовність кроків при побудові моделі ГРІД (рис.1.9):

1) статистичний аналіз і опис характеру потоку вхідних задач;

2) побудова моделі потоку задач, її верифікація и калібрування;

3) аналіз і побудова моделі продуктивності і доступності ГРІД-ресурсів;

4) аналіз і побудова моделі ГРІД на основі структурних компонент (ресурсів), які беруть участь в розробці задач, функціональних зв'язків між ними і потоку вхідних заявок.

Одним з механізмів підвищення відмовостійкості ГРІД-системи є оптимальне (зниження відмов виконання заявок) планування ПЗ між всіма ресурсами, тому розуміння її структури і характеристик є необхідною умовою побудови адекватної моделі системи в цілому.

Можна виділити наступні кроки побудови робочого навантаження:

1) ідентифікація цілей моделювання:

- вибір компонент (workload components) і параметрів (workload parameters) РН, де компонент - базова сутність, формує робоче навантаження (у нашому випадку - задача) і параметр - кількісний показник, що характеризує компонент навантаження (тип завдання, розмір задачі, вимоги до ресурсів і т.д.);

- визначення критерію адекватності моделі;

2) моніторинг і збір даних по вибраним характеристикам;

3) статистичний аналіз отриманих даних:

- попередній аналіз: опис оброблюваних даних: аналізований часовий інтервал, кількість оброблюваних значень, визначення аномалій у робочому навантаженні і прийняття рішення про їх виключення або включення в майбутню модель;

- описова статистика досліджуваних величин: частотний аналіз (аналіз гістограми, підгонка розподілу і його оцінка); обчислення та аналіз показників центру розподілу (мода, медіана, середнє); оцінка розкиду даних у сукупності (стандартне відхилення, дисперсія, коефіцієнти варіації);


Рисунок 1.9 – Методологія аналізу відмовостійкості ГРІД-системи

- динамічний аналіз: аналіз досліджуваної величини у часі, побудова ЧР, визначення тренду, циклічної і сезонної складової.

4) побудова аналітичної або імітаційної моделі потоку робочого навантаження, наприклад: використання Марковських моделей, граф-схем, графів станів і переходів, апарату мереж Петрі і систем масового обслуговування, моделювання потоку задач у вигляді сукупності потоків з різними показниками (наприклад, інтенсивності) і т. д.

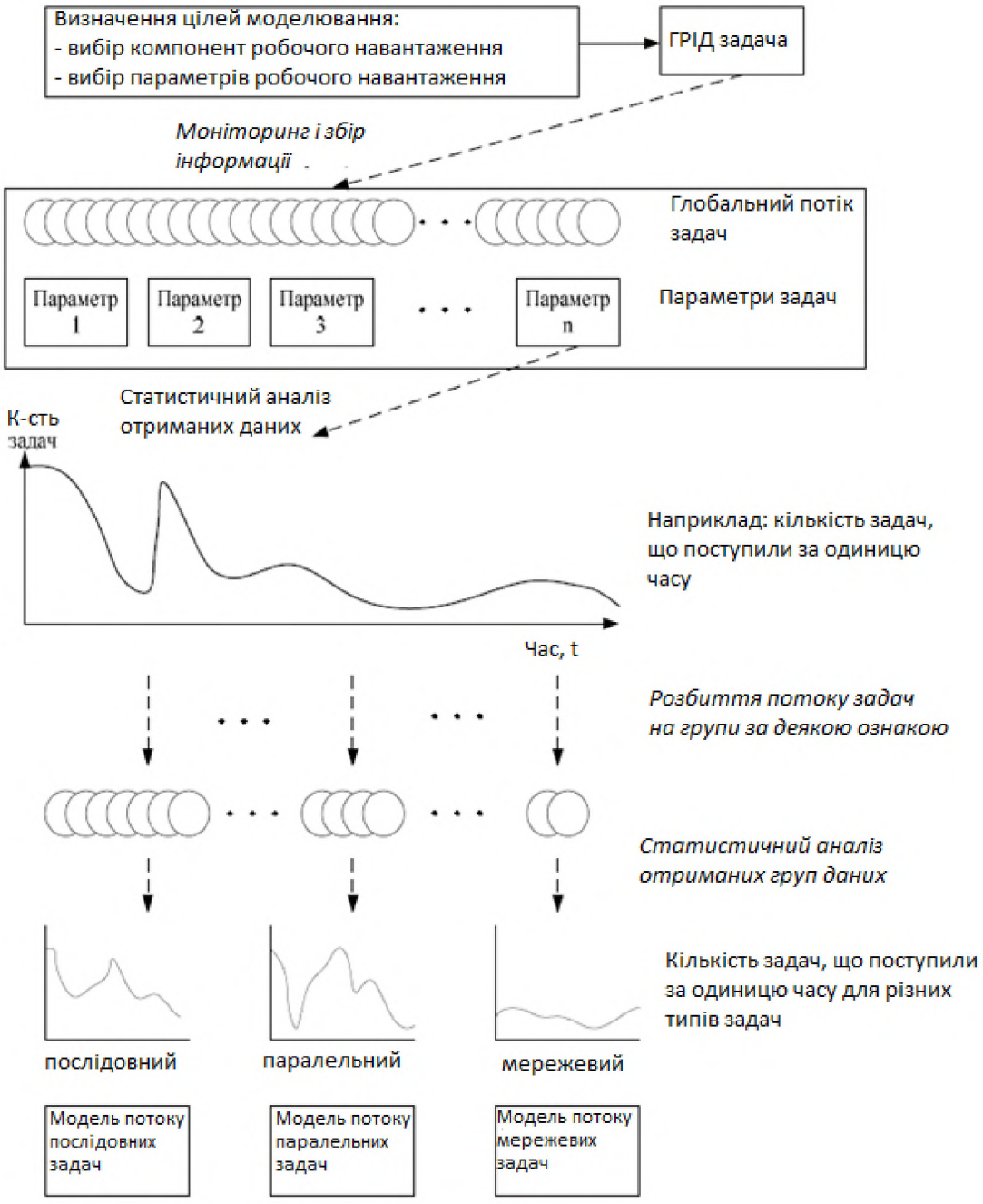



Рисунок 1.10 - Загальна послідовність побудови моделі потоку задач

5) перевірка адекватності і точності моделі може бути виконана на основі деякого критерію, наприклад продуктивності (пропускна здатність системи, час обробки одиниці потоку, навантаження на обслуговуючі вузли і т.д.) тоді точність моделі оцінюється як функція різниці значень отриманих в результаті обробки дійсного потоку РН (real workload) і його моделі.

На рисунку. 1.10 наведена послідовність побудови моделі ПЗ залежно від їх типу (рис. 1.5), де результатом є не загальна модель глобальної РН, а сукупність моделей навантажень, що надходять в систему. Перевагою даного підходу є можливість впливати на процес моделювання безпосередньо на рівні типів задач, що в свою чергу дозволяє спрогнозувати поведінку системи в разі збільшення/зменшення кількості, наприклад паралельних задач або задач послідовного типу, тобто тих, для яких потрібно один процесор.

## 1.6 Статистичний аналіз характеру потоку задач в ГРІД

Статистичний аналіз присвячений вирішенню задачі найкращого підбору, тобто статистичного оцінювання невідомих параметрів, що входять в аналітичний запис моделі, і дослідженню властивостей отриманих оцінок, їх точності. Розв'язання даної задачі повністю обслуговується методами статистичної обробки даних.

Оцінка продуктивності або інших параметрів ГРІД з використанням методів імітаційного моделювання вимагає побудови адекватних моделей потоку вхідних подій і формалізації процесу його обробки.

Складність потоку найкращим способом можна описати на основі емпіричних даних, отриманих як на основі прогону реального ПЗ (отриманого в результаті документування процесу обробки подій), так і на основі його синтетичних моделей (отриманих в результаті ймовірнісно-статистичного аналізу журналу потоку подій).

Статистичний аналіз потоку задач і побудову її моделі можна розбити на три основні етапи:

а) попередній аналіз або первинна обробка даних – у ході первинної статистичної обробки даних зазвичай вирішуються такі задачі:

1) відображення змінних, описаних текстом, в номінальну (з запропонованим числом градацій) або ординальну (порядкову) шкалу;

2) статистичний опис вихідних сукупностей з визначенням меж варіювання змінних;

3) аналіз спостережень, що різко виділяються, пояснення і усунення аномалій у параметрах робочого навантаження (наприклад, негативний час виконання задачі, який в ряді ГРІД систем можна виключити з дослідження, а в інших навпаки - інформувати про мінімальний час виконання задачі);

4) відновлення пропущених спостережень;

5) перевірка статистичної незалежності послідовності спостережень, що складають масив вихідних даних;

6) уніфікація типів змінних, коли за допомогою різних прийомів домагаються уніфікованого запису всіх змінних;

б) описова статистика – даний етап включає:

1) обчислення й аналіз показників центру розподілу (мода, медіана, середнє);

2) частотний аналіз (аналіз гістограми, попередня підгонка розподілу та його оцінка, перевірка на нормальність закону розподілу досліджуваних величин);

3) оцінка розкиду даних у сукупності (стандартне відхилення, дисперсія, коефіцієнти варіації);

4) кластерний аналіз. Класифікація і розбиття компонентів робочого навантаження за деякими загальними ознаками на кілька груп (кластерів). Аналіз отриманих груп з використання методів розглянутих вище;

в) побудова моделі потоку задач – найчастіше даний етап включає наступні кроки:

1) експериментальний аналіз закону розподілу досліджуваної генеральної сукупності (або отриманих груп - кластерів) і параметризація відомостей про природу досліджуваних розподілів;

2) перевірка адекватності отриманої моделі потоку задач з використанням графічних і чисельних методів, наприклад використання критеріїв згоди для перевірки статистичних гіпотез про відповідності ЕФР досліджуваних параметрів робочого навантаження теоретичній (goodness of fit, GoF), тобто чи відповідає отриманий розподіл передбачуваній моделі. Наприклад, час надходження задач у ГРІД розподілений за експоненціальним законом.

### 1.6.1 Розробка дочірнього компоненту WeatherSummary

Результати статистичного аналізу зручно представляти у двох видах – табличному (кількісному) і графічному.

Розглянемо методи і характеристики, що дають нам представлення про кількісні числові характеристики:

1) позначимо множину даних $X = (x_1, x_2, ..., x_n)$. Середнім арифметичним (математичним очікуванням) називається сума всіх чисел ряду, поділена на їх кількість.

$$\bar{x} = \frac{1}{n} \cdot (x_1 + x_2 + ... + x_n), \qquad (1.1)$$

2) мінімум – найменше значення вибірки;

3) максимум – найбільше значення вибірки;

4) мода – значення випадкової величини, що трапляється найчастіше в сукупності спостережень. Іноді трапляється більше аніж одна мода - у такому випадку можна сказати, що сукупність мультимодальна. Із структурних середніх величин лише мода має таку унікальну властивість. Як правило

мультимодальність вказує на те, що набір даних не підпорядковується нормальному розподілу:

$$Mo = a_i + k \cdot \frac{n_i - n_{i-1}}{2 \cdot n_i - n_{i-1} - n_{i+1}} \quad . \qquad (1.2)$$

5) медіана – характеристика, що розділяє впорядкований (за зростанням чи спаданням) ряд експериментальних даних на дві рівні частини. Медіаною називається таке число, що задовольняє наступну умову:

$$Me = \begin{cases} x_{\frac{n+1}{2}} \; при \; n = 2k + 1, \\ x_{\frac{n}{2}+1} + x_{\frac{n}{2}} \; при \; n = 2k.' \end{cases} \qquad (1.3)$$

тобто, ймовірність того, що випадкова величина матиме значення більше або менше за медіану однакова і дорівнює 1/2.

6) дисперсія – міра розсіювання випадкової величини, тобто її відхилення від математичного сподівання:

$$D[X] = M\left[\left|X - M[X]\right|^2\right] = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad (1.4)$$

де $M$ математичне сподівання;

7) стандартне відхилення – це квадратний корінь із дисперсії, що вказує на розсіювання значень випадкової величини відносно її математичного сподівання;

$$\sigma = \sqrt{D[X]} \, . \qquad (1.5)$$

8) коефіцієнт варіації – міра відносного розсіювання випадкової величини, що показує, яку долю середнього значення цієї величини складає її середнє розсіювання:

$$V = \frac{\sigma}{\bar{x}} \, . \qquad (1.6)$$

Коефіцієнт варіації можна обчислити, розділивши стандартне відхилення на середнє арифметичне значення змінної, виражене у відсотках. Результат даного обчислення може потрапляти в інтервал від нуля до нескінченності, зростаючи в міру збільшення варіації ознаки. Якщо отримане значення менше 33,3% – варіація ознаки слабка. Якщо більше – сильна. В останньому випадку

досліджувана сукупність даних є неоднорідною, її середня величина визнається нетиповою, а тому не може бути узагальнюючим показником. Тому для даної сукупності варто застосувати інші показники.

9) асиметрія – міра несиметричності розподілу. Якщо асиметрія (що показує відхилення розподілу від симетричного) істотно відрізняється від 0, то розподіл несиметричний. У симетричного розподілу асиметрія рівна 0. Коефіцієнт асиметрії визначається третім центральним розподілом:

$$A_x = \frac{M\left[\left(X - M[X]\right)^3\right]}{\sigma^3} \cdot 100\%. \qquad (1.7)$$

10) ексцес – міра гостроти піку розподілу випадкової величини. Він показує, на скільки гостру вершину має щільність розподілу у порівнянні з нормальним розподілом. Якщо коефіцієнт ексцесу більше 0, то розподіл має більш гостру вершину, ніж нормальний розподіл, якщо менше 0 – більш плоску:

$$E_x = \frac{M\left[\left(X - M[X]\right)^4\right]}{\sigma^4} - 3. \qquad (1.8)$$

### 1.6.2 Функція розподілу, візуальне представлення кількісних характеристик

Нехай потрібно дослідити сукупність однорідних об’єктів відносно деякої якісної чи кількісної ознаки.

Вибірковою сукупністю або просто вибіркою називають сукупність випадково відібраних об’єктів. Генеральною сукупністю називають сукупність об’єктів, із яких вилучається вибірка.

Об’ємом вибіркової сукупності називають число об’єктів у цій сукупності.

Нехай є вибірка $X = (x_1, x_2, ..., x_k)$, при чому $x_1$ спостерігалося $n_1$ разів, $x_2$ – $n_2$ разів, $x_k$ – $n_k$ раз і $\sum n_i = n$ – об’єм вибірки. Спостережувані значення $x_i$ називають варіантами, а послідовність варіант, записаних у зростаючому

порядку, - варіаційним рядом. Числа спостережень називають частотами, а їх відношення до об'єму вибірки $n_i / n = W$ - відносними частотами.

Статистичним розподілом вибірки називають перелік варіант і відповідних їм частот чи відносних частот. Статистичний розподіл можна задати також у вигляді послідовності інтервалів і відповідних їм частотам (в якості частоти, що відповідає заданому інтервалу, приймають суму частот, що попали в цей інтервал). Під розподілом розуміють відповідність між спостережуваними варіантами і їх частотами чи відносними частотами.

Розподіл ймовірностей – закон, що описує область значення випадкової величини і ймовірність її прийняття. Нехай задано статистичний розподіл частот кількісної ознаки частот *X*. Нехай $n_x$ - число спостережень, при яких спостерігалося значення ознаки, менше за $x$, $n$ - загальна кількість спостережень (об'єм вибірки). Ясно, що відносна частота події $X < x$ рівна $n_x / n$. Якщо $x$ змінюється, то, взагалі кажучи, змінюється і відносна частота, тобто відносна частота $n_x / n$ є функція від $x$. Так, як ця функція знаходиться емпіричним (дослідним) шляхом, її називають емпіричною.

Емпіричною функцією розподілу (функцією розподілу вибірки) називають функцію $F^*(x)$, що визначає для кожного значення $x$ відносну частоту події $X < x$.

Таким чином, за визначенням, $F^*(x) = n_x / n$, де $n_x$ - число варіант, менших за $x$, $n$ - об'єм вибірки. Щоб знайти, наприклад, $F^*(x_2)$, потрібно число варіант, менших за $x_2$, поділити на об'єм вибірки: $F^*(x_2) = n_{x_2} / n$.

На відміну від ЕФР вибірки функцію розподілу $F(x)$ генеральної сукупності називають теоретичною функцією розподілу. Різниця між ЕФР і ТФР заключається в тому, що ТФР $F(x)$ визначає ймовірність події $X < x$, а ЕФР $F^x(x)$ визначає відносну частоту цієї ж події.

Властивості ЕФР:

1) значення ЕФР належать відрізку $[0;1]$;

2) $F^*(x)$ - неспадна функція;

3) якщо $x_1$ – найменша варіанта, то $F^*(x)=0$ при $x \le x_1$; якщо $x_k$ – найбільша варіанта, то $F^*(x)=1$ при $x > x_k$.

Таким чином, ЕФР вибірки служить для оцінки ТФР.

Існує ряд методів представлення функції розподілу:

1) ЕФР також називають кумулятивною функцією розподілу (cumulative distribution function - cdf);

2) функція щільності ймовірності (probability distribution function - pdf) – визначає ймовірність того, що випадкова величина буде знаходитися в околі визначеної точки пропорційна щільності розподілу ймовірності випадкової величини в цій точці:

$$f(t)=dF(t)/dt\,. \qquad (1.9)$$

3) комплементарна інтегральна функція розподілу (complementary cumulative distribution function - ccdf) являється оберненою до ФР:

$$F'(x)=1-F(x). \qquad (1.10)$$

Для візуального представлення кількісних характеристик використовують графічні методи, зокрема полігон частот, гістограму, часовий ряд.

Полігоном частот називають ламану криву, відрізки якої з'єднують точки $(x_1,n_1),(x_2,n_2),...,(x_k,n_k)$.

Для побудови полігона частот на осі абсцис відкладають варіанти $x_i$, а на осі ординат – відповідні їм частоти $n_i$. Точки $(x_i,n_i)$ з'єднуються відрізками прямих і отримуємо полігон частот.

Полігоном відносних частот називають ламану, відрізки якої з'єднують точки $(x_1,W_1),(x_2,W_2),...,(x_k,W_k)$.

Для побудови полігона відносних частот на осі абсцис відкладають варіанти $x_i$, а на осі ординат – відповідні їм відносні частоти $W_i$.

Точки $(x_i, W_i)$ з'єднуються відрізками прями і отримуємо полігон відносних частот. У випадку неперервної ознаки доцільно будувати гістограму.

Для цього інтервал, в якому заключені всі спостережувані значення ознаки, розбивають на декілька частинних інтервалів довжиною $h$ і знаходять для кожного частинного інтервалу $n_i$ - суму частот варіант, що попали в *i*–ий інтервал.

Гістограмою частот називають ступінчату фігуру, що складається із прямокутників, основами яких служать частинні інтервали довжиною $h$, а висоти рівні відношенню $n_i / n$ (щільність частоти). Для побудови гістограми частот на осі абсцис відкладають частинні інтервали, а над ними проводять відрізки, паралельні осі абсцис на відстані $n_i / h$.

Площа *i*–го частинного прямокутника рівна $h \cdot n_i / h = n_i$ - сумі частот варіант *i*–го інтервалу. Звідси слідує, що площа гістограми частот рівна сумі всіх частот, тобто об'єму вибірки.

Часовий ряд – зібраний в різні моменти часу статистичний матеріал по значеннях яких-небуть параметрів (в найпростішому випадку одного) досліджуваного процесу.

Кожна одиниця статистичного матеріалу називається відліком, також допустимо називати його рівнем на вказаний з ним момент часу.

У часовому ряді кожному відліку повинен бути вказаний час виміру чи номер виміру по порядку. Часовий ряд істотно відрізняється від простої вибірки даних, так як при аналізі враховується взаємозв'язок із часом, а не тільки статистична різноманітність і статистичні характеристики вибірки.

Для побудови графіка ЧР потрібно на осі абсцис відкласти шкалу періодів (починаючи з нуля), а на осі ординат - значення ознаки у вибраний період. Графік часового ряду показує, як змінюється статистична величина у часі.

### 1.6.3 Пошук теоретичного розподілу. Метод максимальної ймовірності

Закон теоретичного розподілу можна підібрати, дивлячись на вигляд гістограми чи ccdf. Існує ряд методів для визначення параметрів розподілу:

- метод найменших квадратів (МНК);

- (узагальнений) метод моментів;

- метод максимальної вірогідності (МКВ).

Розглянемо метод максимальної вірогідності. Даний метод був запропонований Рональдом Фішером.

Нехай $X$ – неперервна випадкова величина, яка в результаті $n$ випробувань прийняла значення $x_1, x_2, ..., x_n$. Допустимо, що вид закону розподілу випадкової величини $X$ заданий, але невідомий параметр $\theta$, яким визначається цей закон. Потрібно знайти його точкову оцінку.

Позначимо ймовірність того, що в результаті випробування величина $X$ прийме значення $x_i \left(i = \overline{1,n}\right)$, через $p(x_i;\theta)$.

Функцією вірогідності неперервної випадкової величини $X$ називають функцію аргументу $\theta$:

$$L(x_1, x_2, ..., x_n, \theta) = p(x_1;\theta)\, p(x_2;\theta) ... p(x_n;\theta) \ , \qquad (1.11)$$

де $x_1, x_2, ..., x_n$ – фіксовані числа.

В якості точкової оцінки параметра $\theta$ приймають таке його значення $\theta^* = \theta^*(x_1, x_2, ..., x_n)$, при якому функція вірогідності досягає максимуму.

Оцінку $\theta^*$ називають оцінкою найбільшої вірогідності.

Функції $L$ і $\ln L$ досягають максимуму при одному і тому ж значенні $\theta$, тому замість пошуку максимуму функції $L$ шукають (що більш зручно) максимум функції $\ln L$. Логарифмічною функцією вірогідності називають функцію $\ln L$. Як відомо, точку максимуму функції $\ln L$ аргументу $\theta$ можна шукати, наприклад, так:

1) шукається похідна $\frac{d \ln L}{d\theta}$;

2) прирівнюється похідна до нуля і шукається критична точка – корінь отриманого рівняння (його також називають коренем вірогідності);

3) шукається друга похідна $\frac{d^2 \ln L}{d\theta^2}$: якщо друга похідна при $\theta = \theta^*$ від'ємна, то $\theta^*$ - точка максимуму.

Знайдену точку максимуму $\theta^*$ приймають в якості оцінки найбільшої вірогідності параметра $\theta$.

Метод максимальної вірогідності має ряд переваг: оцінки найбільшої вірогідності, взагалі кажучи, досить точні (але можуть бути зміщеними), розподілені асимптотично нормально (при великих значеннях $n$ приблизно нормальні) і мають найменшу дисперсію у порівнянні з іншими асимптотичними нормальними оцінками; якщо для оцінюваного параметра $\theta$ існує ефективна оцінка $\theta^*$, то рівняння вірогідності має єдиний розв'язок $\theta^*$.

Даний метод найбільш повно використовує дані вибірки про оцінюваний параметр, тому він особливо корисний у випадках із малими вибірками.

### 1.6.4 Пошук надійних інтервалів

Точковою називають оцінку, яка визначається одним числом. При виборі малого об'єму вибірки точкова оцінка може значно відрізнятися від оцінюваного параметра, тобто приводити до грубих похибок. По цій причині при невеликому об'єму вибірки потрібно користуватися інтервальними оцінками.

Інтервальною називають оцінку, яка визначається двома числами – кінцями інтервалу. Інтервальні оцінки дозволяють установити точність і надійність оцінок.

Нехай знайдена за даними вибірки статистична характеристика $\theta^*$ служить оцінкою невідомого параметра $\theta$. Будемо вважати $\theta$ константою (число $\theta$ може бути і випадковою величиною). Ясно, що $\theta^*$ тим точніше визначає параметр $\theta$, чим менша абсолютна величина різниці $|\theta - \theta^*|$. Іншими словами, якщо $\delta > 0$ і $|\theta - \theta^*| < \delta$, то чим менше $\delta$, тим оцінка точніше. Таким чином, додатне число $\delta$ характеризує точність оцінки.

Однак статистичні методи не дозволяють категорично стверджувати, що оцінка $\theta^*$ задовольняє нерівність $|\theta-\theta^*|<\delta$, можна лише говорити про ймовірність $\gamma$, з якою дана нерівність виконується.

Надійністю (надійною ймовірністю) оцінки $\theta$ по $\theta^*$ називають ймовірність $\gamma$, з якою виконується нерівність $\gamma$. Зазвичай надійність оцінки задається наперед, при чому в якості $\gamma$ беруть число, що лежить досить близько біля одиниці. Найбільш часто задають надійність, що рівна 0,95; 0,99 і 0,999.

Нехай ймовірність того, що $|\theta-\theta^*|<\delta$, рівна $\gamma$:

$$P\left[|\theta-\theta^*|<\delta\right]=\gamma. \qquad (1.12)$$

Замінивши нерівність $|\theta-\theta^*|<\delta$ рівносильною їй подвійною нерівністю $-\delta<\theta-\theta^*<\delta$, чи $\theta^*-\delta<\theta<\theta^*+\delta$, маємо

$$P\left[\theta^*-\delta<\theta<\theta^*+\delta\right]=\gamma. \qquad (1.13)$$

Це співвідношення потрібно розуміти так: ймовірність того, що інтервал $(\theta^*-\delta,\theta^*+\delta)$ включає в себе (покриває) невідомий параметр $\theta$, рівна $\gamma$.

Надійним інтервалом називають інтервал $(\theta^*-\delta,\theta^*+\delta)$, який покриває невідомий параметр із заданою надійністю $\gamma$.

Досі ми шукали точкові оцінки числових характеристик генеральної сукупності. Виявляється, що для вибірок невеликого обсягу вони можуть сильно відхилятися від реальних значень, тому далі знайдемо інтервальні оцінки тих самих характеристик.

Нехай кількісна ознака $X$ генеральної сукупності розподілена нормально, при чому середньоквадратичне відхилення $\sigma$ цього розподілу відоме. Потрібно оцінити невідоме математичне сподівання $a$ по вибірковому середньому $\bar{x}$. Потрібно знайти НІ, що покривають параметр $a$ з надійністю $\gamma$ (те саме – рівень значущості $\alpha$=1-$\gamma$). Будемо розглядати вибіркове середнє $\bar{x}$ як випадкову величину $\bar{X}$ і вибіркові значення ознаки $x_1,x_2,...,x_n$ – як однаково розподілені

незалежні випадкові величини $X_1, X_2, ..., X_n$. Іншими словами, математичне очікування кожної із цих величин рівне $a$ і середньоквадратичне відхилення - $\sigma$.

Якщо випадкова величина $X$ розподілена нормально, то вибіркова середня $\overline{X}$ знайдена по незалежними спостереженнями, також розподілена нормально. Параметри розподілу $\overline{X}$ такі:

$$M\left(\overline{X}\right)=a, \quad \sigma\left(\overline{X}\right)=\frac{\sigma}{\sqrt{n}}, \tag{1.14}$$

Потрібно, щоб виконувалося відношення

$$P\left(\left|\overline{X}-a\right|<\delta\right)=\gamma \ , \tag{1.15}$$

де $\gamma$ – задана надійність.

Виконаємо перетворення правої частини:

$$P\left(\left|\overline{X}-a\right|<\delta\right)=2\Phi\left(\frac{\delta}{\sigma}\right), \tag{1.16}$$

Здійснимо заміну $X$ на $\overline{X}$, і $\sigma$ на $\sigma\left(\overline{X}\right)=\frac{\sigma}{\sqrt{n}}$, в результаті чого отримаємо

$$P\left(\left|\overline{X}-a\right|<\delta\right)=2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right)=2\Phi\left(t\right), \tag{1.17}$$

де $t=\frac{\delta\sqrt{n}}{\sigma}$.

Знайшовши із останньої рівності $\delta=\frac{t\sigma}{\sqrt{n}}$, можемо записати

$$P\left(\left|\overline{X}-a\right|<\frac{t\sigma}{\sqrt{n}}\right)=2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right)=2\Phi\left(t\right), \tag{1.18}$$

Прийнявши до уваги те, що ймовірність $P$ задана і рівна $\gamma$, остаточно маємо

$$P\left(\overline{x}-\frac{t\sigma}{\sqrt{n}}<a<\overline{x}+\frac{t\sigma}{\sqrt{n}}\right)=2\Phi\left(t\right)=\gamma, \tag{1.19}$$

Сенс отриманого співвідношення такий: з надійністю $\gamma$ можна стверджувати, що НІ $\left(\overline{x}-\frac{t\sigma}{\sqrt{n}},\overline{x}+\frac{t\sigma}{\sqrt{n}}\right)$ покриває невідомий параметр $a$, точність оцінки $\delta=\frac{t\sigma}{\sqrt{n}}$.

Число $t$ визначається із рівності $2\Phi(t)=\gamma$, або $\Phi(t)=\frac{\gamma}{2}$, де аргумент $t$ знаходиться за функцією Лапласа. Нехай кількісна ознака $X$ ГС розподілена нормально, при чому середньоквадратичне відхилення $\sigma$ невідоме. Потрібно оцінити невідоме математичне сподівання $a$ за допомогою НІ.

По даним вибірки можна побудувати випадкову величину (її можливі значення позначимо через *t*):

$$T=\frac{\overline{X}-a}{S/\sqrt{n}}, \tag{1.20}$$

яка має розподіл Стьюдента із $k=n-1$ степенями свободи. Тут $\overline{X}$ - вибіркове середнє; $S$ – виправлене середньоквадратичне відхилення; $n$ – об'єм вибірки.

Щільність розподілу Стьюдента:

$$S(t,n)=B_n\left[1+\frac{t^2}{n-1}\right]^{-n/2}, \tag{1.21}$$

де $B_n=\frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma\left((n-1)/2\right)}$,

де $\Gamma$ – гамма-функція.

Розподіл Стьюдента визначається параметром $n$ - об'ємом вибірки (або числом степенів свободи $k=n-1$) і не залежить від невідомих параметрів $a$ і $\sigma$; ця є значною особливістю даного розподілу.

Оскільки $S(t, n)$ – парна функція від $t$, ймовірність здійснення нерівності $\left|\frac{\overline{X}-a}{S/\sqrt{n}}\right|<\gamma$ визначається так:

$$P\left(\left|\frac{\overline{X}-a}{S/\sqrt{n}}\right|<t_\gamma\right)=2\int_0^{t_\gamma}S(t,n)dt=\gamma. \tag{1.22}$$

Замінивши нерівність у круглих дужках рівносильною їй подвійною нерівністю, отримаємо

$$P\left(\overline{X}-\frac{t_\gamma S}{\sqrt{n}}<a<\overline{X}+\frac{t_\gamma S}{\sqrt{n}}\right)=\gamma. \tag{1.23}$$

Таким чином, використовуючи розподіл Стьюдента, ми знайшли НІ $\left(\bar{x}-\frac{t_\gamma S}{\sqrt{n}}, \bar{x}+\frac{t_\gamma S}{\sqrt{n}}\right)$, що покриває невідомий параметр $a$ з надійністю $\gamma$. Тут випадкові величини $\overline{X}$ і $S$ були замінені невипадковими величинами $\bar{x}$ і $s$, що були знайдені за вибіркою.

Нехай потрібно оцінити невідоме генеральне середньоквадратичне відхилення $\sigma$ по виправленому вибірковому середньоквадратичному відхиленні $s$. Знайдемо НІ, що покривають параметр $\sigma$ із заданою надійністю $\gamma$.

Потрібно, щоб виконувалося співвідношення $P(|\sigma - s| < \delta) = \gamma$, або $P(s-\delta < \sigma < s+\delta) = \gamma$.

Перетворимо подвійну нерівність $s-\delta<\sigma<s+\delta$ в рівносильну нерівність $s\left(1-\frac{\delta}{s}\right)<\sigma<s\left(1+\frac{\delta}{s}\right)$. Поклавши $\frac{\delta}{s}=q$, отримаємо:

$$s(1-q)<\sigma<s(1+q). \tag{1.24}$$

Знайдемо $q$. Для цього введемо випадкову величину «хі»:

$$\chi=(S/\sigma)\sqrt{n-1}, \tag{1.25}$$

де $n$ – об'єм вибірки.

Величина $\frac{S^2(n-1)}{\sigma^2}$ розподілена за законом $\chi^2$ із $n$-$1$ степенями свободи, тому квадратний корінь із неї позначають через $\chi$.Щільність розподілу $\chi$ має вид:

$$R(\chi,n)=\frac{\chi^{n-2}e^{-\chi^2/2}}{2^{(n-3)/2}\Gamma\left(\frac{n-1}{2}\right)}. \tag{1.26}$$

Цей розподіл не залежить від оцінюваного параметра $\sigma$, а залежить лише від об'єму вибірки $n$. Припускаючи, що $q<1$, перепишемо (*) так:

$$\frac{1}{s(1+q)}<\frac{1}{\sigma}<\frac{1}{s(1-q)}. \tag{1.27}$$

Помноживши всі члени нерівності на $S\sqrt{n-1}$, отримаємо:

$$\frac{\sqrt{n-1}}{(1+q)}<\frac{S}{\sigma}<\frac{\sqrt{n-1}}{(1-q)}, \tag{1.28}$$

або

$$\frac{\sqrt{n-1}}{(1+q)} < \chi < \frac{\sqrt{n-1}}{(1-q)}. \qquad (1.29)$$

Ймовірність того, що нерівність, а отже, і рівносильна їй нерівність (*) буде здійснено, рівна

$$\int_{\sqrt{n-1}/(1+q)}^{\sqrt{n-1}/(1-q)} R(\chi, n) d\chi = \gamma. \qquad (1.30)$$

Обчисливши по вибірці $s$ і знайшовши $q$ за таблицями розподілу «хі», отримаємо шуканий НІ, що покриває $\sigma$ із заданою ймовірністю $\gamma$.

### 1.6.5 Перевірка адекватності теоретичного розподілу

Використання функції розподілу дозволяє однозначно задати емпіричний закон розподілу досліджуваної величини і оцінити його відповідність теоретичному закону за допомогою критеріїв згоди.

Критеріями згоди називають статистичні критерії, призначені для перевірки згоди досліджуваних даних і теоретичної моделі. Краще всього застосовувати критерії, якщо спостереження представляють випадкову вибірку. Теоретична модель в цьому випадку описує закон розподілу.

Існує ряд критеріїв згоди, зокрема:

- омега-квадрат;
- хі-квадрат;
- Колмогорова;
- Колмогорова-Смірнова;
- Андерсона-Дарлінга.

Розглянемо критерій згоди Колмогорова. Нехай маємо вибірку об'ємом *n*. Позначимо істинну функцію розподілу, якій підкоряється спостереження, *G(x)*, ЕФР $F_n(x)$, а ТФР *F(x)*. Тоді введемо гіпотезу *H0*, яка буде стверджувати, що істинна функція розподілу є *F(x)* і вона записується у виді: $H0: G(x) = F(x)$. Також введемо альтернативну гіпотезу $H1: G(x) \neq F(x)$.

Якщо нульова гіпотеза справджується, то $F_n$ і $F$ повинні проявляти визначену схожість і різниця між ними повинна спадати із збільшенням $n$. Внаслідок теореми Бернуллі $F_n(x) \to F(x)$ при n $\to \infty$ . Для кількісного вираження збіжності функцій використовують $F_n$ и $F$ різні способи.

Для вираження збіжності функцій використаємо наступну метрику:

$$D_n = \sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right|. \qquad (1.31)$$

Статистику $D_n$ називають статистикою Колмогорова.

Очевидно, що – $D_n$ випадкова величина, оскільки її значення залежить від $F_n$. Якщо гіпотеза *H0* справедлива і $n \to \infty$, то $F_n(x) \to F(x)$ при будь-якому $x$. при цих умовах $D_n \to 0$. Якщо ж справджується альтернативна гіпотеза *H1*, то $F_n \to G$ і $G \neq F$, а тому:

$$\sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right| \to \sup_x \left|G(x) - F(x)\right|. \qquad (1.32)$$

Остання величина додатна, так як $G$ не співпадає із $F$. Така відмінність у поведінці $D_n$ у залежності від того, справджується гіпотеза *H0* чи ні, дозволяє використовувати статистику $D_n$ як статистику для перевірки *H0*.

Ключовою властивістю $D_n$ заключається в тому, що якщо $G=F$, тобто гіпотетичний закон вказано правильно, то закон розподілу статистики $D_n$ буде одним і тим же для всіх неперервних функцій $G$. Він залежить тільки від об'єму вибірки $n$.

Доведення даного факту засновано на тому, що статистика не змінює свого значення при монотонних перетвореннях осі $x$. Таким перетворенням будь-який неперервний розподіл $G$ можна перетворити в рівномірний на відрізку *[0, 1]*. При цьому $F_n(x)$ перейде у функцію розподілу вибірки із цього рівномірного розподілу.

При малих $n$ для статистики $D_n$ при гіпотезі *H0* складені таблиці процентних точок. При великих $n$ розподіл $D_n$ (при гіпотезі *H0*) вказує знайдена у 1933р. А. Н. Колмогоровим гранична теорема. У ній говориться про статистику

$\sqrt{n}D$. Теорема Колмогорова стверджує, що при справедливості *H0* і якщо *G* неперервна, то

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n < \lambda\right) = +2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2z^2}. \qquad (1.33)$$

Задаючи рівень значущості $\alpha$ із співвідношення $P(\lambda)=\alpha$ можна знайти відповідне критичне значення $\lambda_\alpha$.

Наведемо у табл. 1.1 критичне значення $\lambda_\alpha$ КК для деяких $\alpha$:

Таблиця 1.1 – Критичні значення $\lambda_\alpha$ критерія Колмогорова

| Рівень значущості $\alpha$ | 0,4 | 0,3 | 0,2 | 0,1 | 0,05 | 0,025 | 0,01 | 0,005 | 0,001 | 0,0005 |
|---|---|---|---|---|---|---|---|---|---|---|
| Критичне значення $\lambda_\alpha$ | 0,89 | 0,97 | 1,07 | 1,22 | 1,36 | 1,48 | 1,63 | 1,73 | 1,95 | 2,03 |

Даний критерій має правосторонню критичну область.

# РОЗДІЛ 2
# РОЗРОБКА БЛОК-СХЕМИ АЛГОРИТМУ

## 2.1 Опис прикладу виконання лабораторної роботи

Розглянемо хід виконання лабораторної роботи для побудови моделі потоку задач у ГРІД.

Наведемо план виконання роботи:

1) на основі даних про інтенсивність потоку задач в місяць (од/год) побудувати ряд розподілу, дати його графічне і табличне представлення (побудувавши на графіку гістограму, полігон частот);

2) знайти основні числові характеристики; оцінити степінь однорідності елементів сукупності;

3) побудувати графіки функції щільності ймовірності і досліджуваних вибірок, а також ccdf;

4) використовуючи середовище Matlab вибрати ряд кандидатів теоретичних розподілів, що найбільше підходять для досліджуваних даних, і оцінити оцінки для невідомих параметрів розподілу, використовуючи метод максимальної вірогідності, що реалізований у виді функції *mle* в пакеті Matlab;

5) побудувати НІ для рівня довіри *1-0,05* для невідомих параметрів розподілу;

6) побудувати і видвинути гіпотезу за КК з критичним рівнем *0,05*; оцінити адекватність вибраних моделей.

В додатку Б наведена вибірка, що показує інтенсивність потоку задач в місяць.

Вирішимо питання, чи доцільно складати варіаційний ряд. Із цією метою знайдемо мінімальні та максимальні елементи вибірки. Маємо: $x_{min} = 32$, $x_{max} = 3735$. Розмах вибірки $x_{max}-x_{min}+1 = 3735-32+1 = 3704$ досить великий, кількість елементів у вибірці 80, тому слід будувати інтервальний варіаційний ряд.

Обчислимо довжину інтервалу для ряду:

$$k_0 = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \ln(n)} = \frac{3735 - 32}{1 + 3{,}322 \cdot \ln(80)} = 238{,}03.$$

Заокруглюємо $k_0$ до найближчого цілого тобто $k_0 = 238$.

Виберемо початок першого інтервалу, його треба вибрати так, щоб перша й остання варіанти не припадали на початок першого і кінець останнього інтервалу відповідно.

Знайдемо перший інтервал:

$$(a_1; b_1) = \left[ x_{\min} - \frac{k}{2}; x_{\min} + \frac{k}{2} \right) = \left[ 32 - \frac{238}{2}; 32 + \frac{238}{2} \right) = [-87; 151).$$

Для наступних інтервалів маємо наступну формулу:

$$(a_i; b_i) = \left[ x_i - \frac{k}{2}; x_i + \frac{k}{2} \right).$$

Всього кількість інтервалів: *m=17*. Для отриманих інтервалів запишемо відносні і накопичені частоти, а також величину *h* для побудови графіка гістограми. Отримані дані наведемо в табл. 1.1.

Знайдемо числові характеристики для вибірки.

Для пошуку вибіркового середнього (математичного сподівання) скористаємося формулою (1.1):

$$\bar{x} = \frac{1}{80} \cdot (x_1 + x_2 + \ldots + x_n) = 712{,}01.$$

Для пошуку вибіркової дисперсії скористаємося формулою (1.4):

$$D[X] = M\left[ \left| X - M[X] \right|^2 \right] = \frac{1}{80} \sum_{i=1}^{n} (x_i - 712{,}01)^2 = 563470{,}34.$$

Для знаходження середньоквадратичного відхилення обчислимо квадратний корінь із дисперсії (формула 1.5):

$$\sigma = \sqrt{D[X]} = \sqrt{563470{,}34} = 750{,}65.$$

Таблиця 2.2 – Інтервальний варіаційний ряд

| i | $a_i$ | $b_i$ | $x_i$ | $n_i$ | $n_i/n$ | $n_i^o$ | $n_i^o/n$ | $n_i/(nk)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | -87,00 | 151,00 | 32 | 15 | 0,1875 | 15 | 0,1875 | 0,000788 |
| 2 | 151,00 | 389,00 | 270 | 23 | 0,2875 | 38 | 0,475 | 0,001208 |
| 3 | 389,00 | 627,00 | 508 | 10 | 0,125 | 48 | 0,6 | 0,000525 |

*Продовження таблиці 2.2.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | 627,00 | 865,00 | 746 | 12 | 0,15 | 60 | 0,75 | 0,00063 |
| 5 | 865,00 | 1103,00 | 984 | 5 | 0,0625 | 65 | 0,8125 | 0,000263 |
| 6 | 1103,00 | 1341,00 | 1222 | 2 | 0,025 | 67 | 0,8375 | 0,000105 |
| 7 | 1341,00 | 1579,00 | 1460 | 2 | 0,025 | 69 | 0,8625 | 0,000105 |
| 8 | 1579,00 | 1817,00 | 1698 | 2 | 0,025 | 71 | 0,8875 | 0,000105 |
| 9 | 1817,00 | 2055,00 | 1936 | 2 | 0,025 | 73 | 0,9125 | 0,000105 |
| 10 | 2055,00 | 2293,00 | 2174 | 2 | 0,025 | 75 | 0,9375 | 0,000105 |
| 11 | 2293,00 | 2531,00 | 2412 | 2 | 0,025 | 77 | 0,9625 | 0,000105 |
| 12 | 2531,00 | 2769,00 | 2650 | 1 | 0,0125 | 78 | 0,975 | 5,25E-05 |
| 13 | 2769,00 | 3007,00 | 2888 | 1 | 0,0125 | 79 | 0,9875 | 5,25E-05 |
| 14 | 3007,00 | 3245,00 | 3126 | 0 | 0 | 79 | 0,9875 | 0 |
| 15 | 3245,00 | 3483,00 | 3364 | 0 | 0 | 79 | 0,9875 | 0 |
| 16 | 3483,00 | 3721,00 | 3602 | 0 | 0 | 79 | 0,9875 | 0 |
| 17 | 3721,00 | 3959,00 | 3840 | 1 | 0,0125 | 80 | 1 | 5,25E-05 |

Для пошуку коефіцієнта варіації розділимо стандартне відхилення на вибіркове середнє (формула 1.6):

$$v = \frac{MX}{\sigma_x} = \frac{712,01}{750,65} = 0,95 .$$

Для пошуку моди скористаємося формулою (1.2). Найбільше варіант містить другий інтервал $(a_2, b_2) = (151, 389)$ - 23:

$$Mo = a_i + k \cdot \frac{n_i - n_{i-1}}{2 \cdot n_i - n_{i-1} - n_{i+1}} = 151 + 238 \cdot \frac{38 - 15}{2 \cdot 38 - 15 - 10} = 297,46 .$$

Медіана *Me* – серединний елемент вибірки, тобто той, справа і зліва від якого в упорядкованій за неспаданням вибірці стоїть однакова кількість елементів. Так, у нашому випадку це буде середина між n/2=40-им та n/2+1=41-им елементами вибірки. За стовпчиком накопичених частот знайдемо варіанти, на які попадають 40-ий та 41-ий елементи вибірки. Це буде варіанта $x_3 = 508$.

Для пошуку коефіцієнта асиметрії скористаємося формулою 1.7:

$$A_x = \frac{M\left[\left(X - M[X]\right)^3\right]}{\sigma^3} = 1,79 .$$

Для пошуку коефіцієнту ексцесу скористаємося формулою 1.8:

$$E_x = \frac{M\left[\left(X - M[X]\right)^4\right]}{\sigma^4} - 3 \approx 3,03 .$$

Міра розсіву задач досить висока (дисперсія рівна 563470,34), тому використовувати середнє значення 712,01 не є доцільно для побудови моделі.

Високий коефіцієнт середньоквадратичного відхилення вказує на значне розсіювання інтенсивності потоку задач від математичного сподівання.

Коефіцієнт варіації, який дорівнює 1,05, вказує на досить високе розсіювання випадкової величини (значні коливання).

Значення коефіцієнта асиметрії ($A=1,79$) вказує на те, що розподіл має довгий правий хвіст. Додатній коефіцієнт ексцесу вказує на те, що пік розподілу загострений. Для візуального представлення кількісних характеристик використаємо графічні методи. Побудуємо полігон відносних частот. На осі $Ox$ відкладемо варіанти $x_i$, на осі на осі $Oy$ – відносні частоти $n_i / n = W$. Полігон відносин частот зображений на рисунку 2.11.

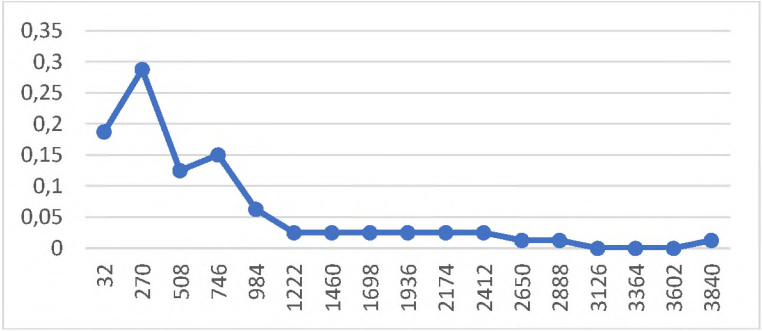



Рисунок 2.11– Полігон відносних частот

Побудуємо гістограму. На осі $Ox$ відкладемо варіанти $x_i$, на осі $Oy$ – висоти $h = n_i / (nk)$. Гістограма зображена на рисунку 2.12.

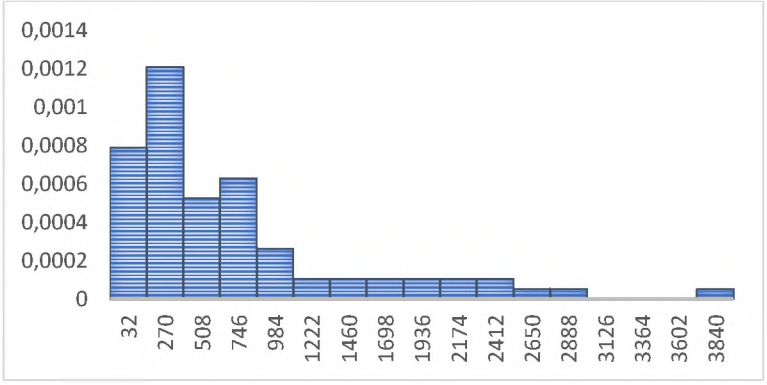



Рисунок 2.12 – Гістограма розподілу частот шуканих даних

Побудуємо графік часового ряду, який показує, як змінюється статистична величина у часі. Для побудови графіка часового ряду потрібно на осі абсцис відкласти шкалу періодів (починаючи з нуля), а на осі ординат - значення ознаки у вибраний період. Графік часового ряду зображений на рисунку 2.13.

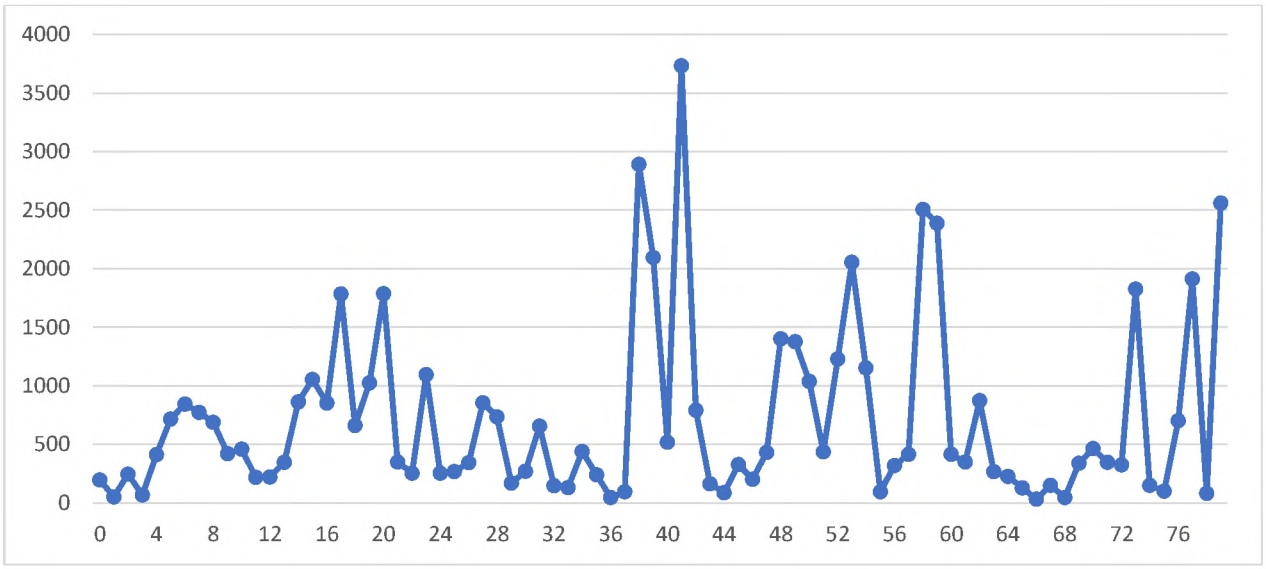



Рисунок 2.13 – Часовий ряд шуканих даних

Побудуємо графік щільності ймовірності. Скористаємося формулою (1.9) –на осі $Ox$ відкладемо варіанти $x_i$, на осі $Oy$ – накопичені частоти $n_i^0 / n$. Графік щільності ймовірності зображений на рисунку 2.14.

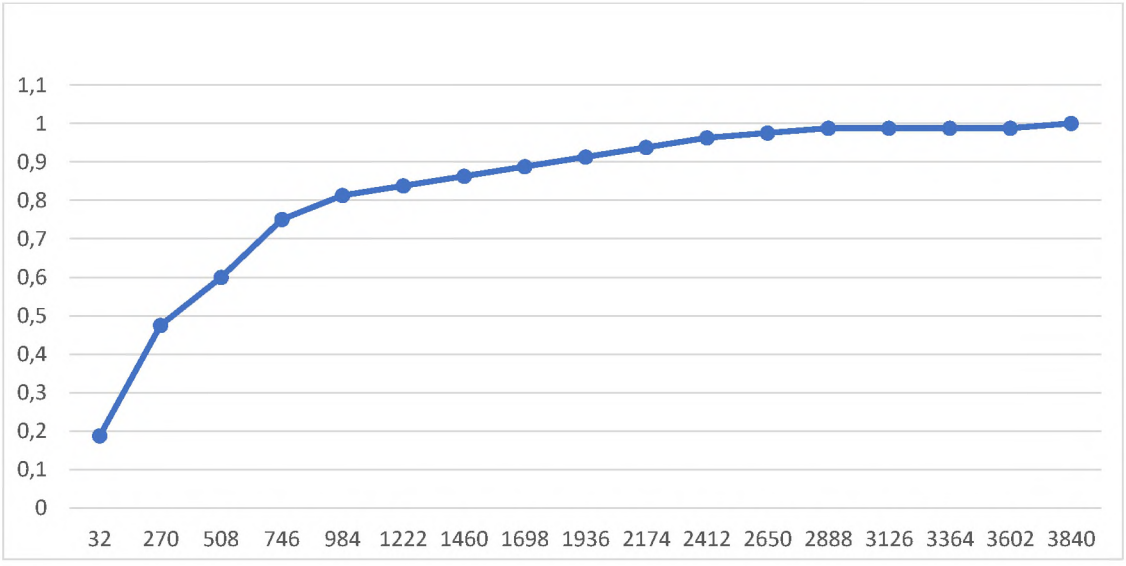



Рисунок 2.14 – Графік щільності ймовірності досліджуваних даних

Побудуємо графік ccdf для емпіричного розподілу – функції, що є оберненою до функції щільності ймовірності (формула (1.9)). Графік ccdf зображений на рисунку 2.15.

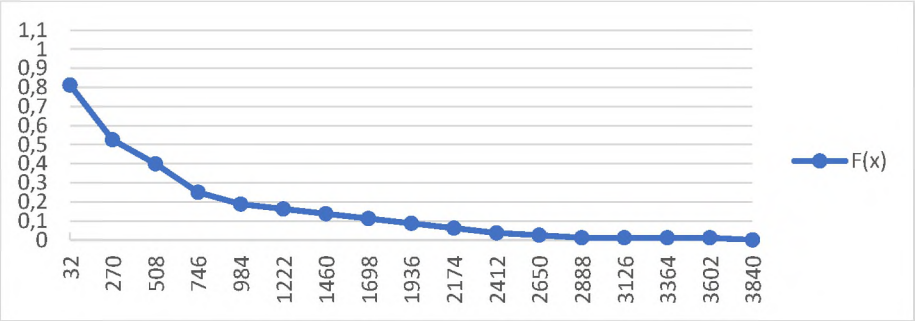



Рисунок 2.15 – Графік ccdf емпіричного розподілу

Ми знайшли тільки точкові оцінки, тобто оцінки, які визначаються одним числом. Звичайно, при користуванні такими оцінками ми не можемо вказати їх точність, тобто наскільки вони відхиляються від істинних значень параметрів. Зрозуміло, що точкові оцінки залежать від обсягу вибірки. Зокрема, якщо обсяг вибірки малий, то точкова оцінка може суттєво відрізнятися від оцінюваного параметра. Тому зручніше користуватися інтервальними оцінками, тобто такими оцінками, які визначаються двома числами - кінцями інтервалу.

Знайдемо НІ для математичного сподівання *MX*, дисперсії *DX*, середньоквадратичного відхилення $\sigma$ з рівнем довіри $\gamma = 1 - 0{,}05$.

Знайдемо НІ для математичного сподівання. Оскільки дисперсія невідома, використаємо формулу (2.34):

$$\Delta = \frac{t \cdot S}{\sqrt{n}}, \qquad (2.34)$$

де *t* шукаємо за таблицею Стьюдента з ступенями свободи який дорівнює $n-1$.

$$t = T_{кр}^{двост}(\alpha; n-1) = T_{кр}^{двост}(0{,}05; 80-1) = 1{,}99.$$

Таким чином, довжина надійного півінтервалу:

$$\Delta = \frac{t \cdot S}{\sqrt{n}} = \frac{1{,}99 \cdot 755{,}38}{\sqrt{80}} = 168{,}1.$$

Надійний інтервал:

$$MX \in (\bar{x} - \Delta; \bar{x} + \Delta) = (712,01 - 168,1; 712,01 + 168,1) = (543,91; 880,11).$$

Знайдемо НІ для дисперсії. Кінці інтервалів знаходяться наступним чином:

$$Dx \in \left( \frac{n-1}{u_2} \cdot S^2; \frac{n-1}{u_1} \cdot S^2 \right),$$

де $u_1 = X^2_{кр}\left(1 - \frac{\alpha}{2}; n-1\right)$;

$$u_2 = X^2_{кр}\left(\frac{\alpha}{2}; n-1\right).$$

Величини $u_1$, $u_2$ визначаються за таблицею критичних точок розподілу Пірсона для n-1-го степеня свободи. таким чином:

$$u_1 = \chi^2_{кр}\left(1 - \frac{\alpha}{2}; n-1\right) = \chi^2_{кр}\left(1 - \frac{0,05}{2}; 80-1\right) = 56,31;$$

$$u_2 = \chi^2_{кр}\left(\frac{\alpha}{2}; n-1\right) = \chi^2_{кр}\left(\frac{0,05}{2}; 80-1\right) = 105,47.$$

Надійний інтервал:

$$DX \in \left( \frac{n-1}{u_2} \cdot S^2; \frac{n-1}{u_1} \cdot S^2 \right) = \left( 79 \cdot \frac{570602,87}{105,47}; 79 \cdot \frac{570602,87}{56,31} \right) = (427386,48; 800541,74).$$

НІ для середньоквадратичного відхилення визначається наступним чином:

$$\sigma_x \in \left( \sqrt{\frac{n-1}{u_2} \cdot S^2}; \sqrt{\frac{n-1}{u_1} \cdot S^2} \right) = \left( \sqrt{79 \cdot \frac{570602,87}{105,47}}; \sqrt{79 \cdot \frac{570602,87}{56,31}} \right) = (653,75; 894,73).$$

Закон теоретичного розподілу можна підібрати, дивлячись на вигляд гістограми чи ccdf.

Якщо розглянути графік ccdf (рис. 2.15), можна побачити, що для інтенсивності потоку задач можна використати наступні теоретичні розподіли:

- гамма розподіл;
- розподіл Вейбула;
- експоненціальний розподіл.

Гама розподіл в теорії ймовірностей — це двопараметричне сімейство абсолютно неперервних розподілів. Він складається з параметрів θ і k. Якщо k - ціле тоді розподіл показує суму k незалежних експоненціально розподілених

випадкових величин, кожна з яких приймає значення θ. Якщо параметр приймає ціле значення, то такий гамма-розподіл також називається розподілом Ерланга. Функція щільності гамма-розподілу має наступний вигляд:

$$f(x)=\begin{cases} x^{k-1}\dfrac{e^{-x/\theta}}{\theta^{k}\Gamma(k)}, x\geq 0, \\ 0, x<0. \end{cases} \tag{2.35}$$

де функція $\Gamma(k)$ має вигляд:

$$\Gamma(k)=\int_{0}^{\infty}x^{k-1}e^{-x}dx, \tag{2.36}$$

і має наступні властивості:

1) $\Gamma(k)=(k-1)\Gamma(k-1)$;

2) $\Gamma(0,5)=\sqrt{\pi}$.

Якщо константи $k,\theta>0$, тоді кажуть, що випадкова величина $X$ має гамма-розподіл з параметрами $k$ і $\theta$, записують так – $X\sim\Gamma(k,\theta)$. Параметр $k$ називають параметром форми, а $\theta$ – параметром масштабу. Розподіл Вейбула в теорії ймовірностей – двопараметричне сімейство неперервних кривих. Названий на честь Валоді Вейбула (англ. Waloddi Weibull), котрий навів детальне описання розподілу в 1951 році. Функція щільності розподілу Вейбула має вигляд:

$$f(x,\lambda,k)=\begin{cases} \dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^{k}}, x\geq 0, \\ 0, x<0. \end{cases}. \tag{2.37}$$

Експоненціальний розподіл – абсолютно неперервний розподіл, що моделює час між двома послідовними завершеннями однієї і тієї ж події.

Випадкова величина має експоненціальний розподіл з параметром $\lambda>0$, якщо щільність розподілу задається наступним видом:

$$f(x)=\begin{cases} \lambda e^{-\lambda x}, x\geq 0, \\ 0 \quad\ , x<0. \end{cases}. \tag{2.38}$$

Використовуючи середовище Matlab знайдемо значення невідомих параметрів описаних вище розподілів за допомогою методу максимальної вірогідності, що реалізований у вигляді функції *mle* .

Таблиця 2.3 – Значення невідомих параметрів для розподілів Вейбула, гамма, експоненціального

| Розподіл | Параметри |
|---|---|
| Гамма | shape = 1,0873<br>scale = 654,84 |
| Вейбула | shape = 1,0125<br>scale = 716 |
| Експоненціальний | mean = 712,0125 |

Для перевірки правильності підбору теоретичного розподілу використаємо КК.

Даний метод дозволить перевірити, наскільки точно досліджувані дані описуються отриманими теоретичними законами розподілу.

Введемо нульову й альтернативну гіпотези.

$H0 = \{F_e(x) = F_d(x)\}$;

$H1 = \{F_e(x) \neq F_d(x)\}$,

де $F_e(x)$ – ЕФР, $F_d(x)$ - передбачувана ТФР із параметрами, визначеними у таблиці 2.4.

.Необхідно перевірити нульову гіпотезу для рівня значущості $\alpha = 0,05$ .Перевіримо адекватність гамма розподілу. Знайдемо значення ccdf для емпіричного і теоретичного законів, також знайдемо абсолютне значення різниці між отриманими значеннями. Отримані дані наведемо у таблиці 2.4.

Таблиця 2.4 – КК для гамма-розподілу:

| $x_i$ | $F_e(x)$ | $1-F_e(x)$ | $F_d(x)$ | $1-F_d(x)$ | D |
|---|---|---|---|---|---|
| 32 | 0,18750 | 0,81250 | 0,0352 | 0,9648 | 0,1523 |
| 270 | 0,47500 | 0,52500 | 0,2980 | 0,7020 | 0,1770 |
| 508 | 0,60000 | 0,40000 | 0,4990 | 0,5010 | 0,1010 |
| 746 | 0,75000 | 0,25000 | 0,6449 | 0,3551 | 0,1051 |
| 984 | 0,81250 | 0,18750 | 0,7492 | 0,2508 | 0,0633 |
| 1222 | 0,83750 | 0,16250 | 0,8234 | 0,1766 | 0,0141 |

*Продовження таблиці 2.4.*

| | | | | | |
|---|---|---|---|---|---|
| 1936 | 0,91250 | 0,08750 | 0,9388 | 0,0612 | 0,0263 |
| 2174 | 0,93750 | 0,06250 | 0,9571 | 0,0429 | 0,0196 |
| 2412 | 0,96250 | 0,03750 | 0,9700 | 0,0300 | 0,0075 |
| 2650 | 0,97500 | 0,02500 | 0,9790 | 0,0210 | 0,0040 |
| 2888 | 0,98750 | 0,01250 | 0,9853 | 0,0147 | 0,0022 |

| | | | | | |
|---|---|---|---|---|---|
| 3126 | 0,98750 | 0,01250 | 0,9897 | 0,0103 | 0,0022 |
| 3364 | 0,98750 | 0,01250 | 0,9928 | 0,0072 | 0,0053 |
| 3602 | 0,98750 | 0,01250 | 0,9950 | 0,0050 | 0,0075 |
| 3840 | 1,00000 | 0,00000 | 0,9965 | 0,0035 | 0,0035 |

Знайдемо міру розходження між емпіричним і теоретичним розподілами:

$$\lambda = D_{\max}\sqrt{n} = D_2\sqrt{80} = 0{,}177 \cdot \sqrt{80} = 1{,}58 .$$

Знайдемо критичне значення для КК за табл. 1.1: $\lambda_\alpha = 1{,}36$.

Оскільки $\lambda > \lambda_\alpha$, звідси слідує, що з ймовірністю *99,5%* ми відхиляємо гіпотезу *H0* і приймаємо гіпотезу *H1*.

Перевіримо адекватність розподілу Вейбула. У табл. 2.5 наведемо значення ccdf для емпіричного і теоретичного законів, також знайдемо абсолютне значення різниці між отриманими значеннями.

Таблиця 2.5 – КК для розподілу Вейбула:

| $x_i$ | $1-F_e(x)$ | $F_e(x)$ | $F_d(x)$ | $1-F_d(x)$ | D |
|---|---|---|---|---|---|
| 32 | 0,18750 | 0,81250 | 0,0421 | 0,9579 | 0,1454 |
| 270 | 0,47500 | 0,52500 | 0,3110 | 0,6890 | 0,1640 |
| 508 | 0,60000 | 0,40000 | 0,5066 | 0,4934 | 0,0934 |
| 746 | 0,75000 | 0,25000 | 0,6474 | 0,3526 | 0,1026 |
| 984 | 0,81250 | 0,18750 | 0,7484 | 0,2516 | 0,0641 |
| 1222 | 0,83750 | 0,16250 | 0,8206 | 0,1794 | 0,0169 |
| 1460 | 0,86250 | 0,13750 | 0,8722 | 0,1278 | 0,0097 |
| 1698 | 0,88750 | 0,11250 | 0,9090 | 0,0910 | 0,0215 |
| 1936 | 0,91250 | 0,08750 | 0,9353 | 0,0647 | 0,0228 |
| 2174 | 0,93750 | 0,06250 | 0,9540 | 0,0460 | 0,0165 |
| 2412 | 0,96250 | 0,03750 | 0,9673 | 0,0327 | 0,0048 |
| 2650 | 0,97500 | 0,02500 | 0,9768 | 0,0232 | 0,0018 |
| 2888 | 0,98750 | 0,01250 | 0,9835 | 0,0165 | 0,0040 |
| 3126 | 0,98750 | 0,01250 | 0,9883 | 0,0117 | 0,0008 |
| 3364 | 0,98750 | 0,01250 | 0,9917 | 0,0083 | 0,0042 |
| 3602 | 0,98750 | 0,01250 | 0,9941 | 0,0059 | 0,0066 |
| 3840 | 1,00000 | 0,00000 | 0,9958 | 0,0042 | 0,0042 |

Знайдемо міру розходження між емпіричним і теоретичним розподілами:

$$\lambda = D_{\max}\sqrt{n} = D_2\sqrt{80} = 0{,}142 \cdot \sqrt{80} = 1{,}27 .$$

Знайдемо критичне значення для КК за таблицею1. 1: $\lambda_\alpha = 1{,}36$.

Оскільки $\lambda < \lambda_\alpha$, звідси слідує, що з ймовірністю *99,5%* ми приймаємо гіпотезу *H0*.

Перевіримо адекватність експоненціального розподілу. У таблиці 2.6 наведемо значення ccdf для емпіричного і теоретичного законів, також знайдемо абсолютне значення різниці між отриманими значеннями.

На рисунку 2.16 покажемо графіки ccdf емпіричної функції та відповідні їй теоретичні функції розподілу.

Таблиця 2.6 – КК для експоненціального розподілу:

| $x_i$ | $F_e(x)$ | 1-Fe(x) | $F_d(x)$ | $1-F_d(x)$ | D |
|---|---|---|---|---|---|
| 32 | 0,18750 | 0,81250 | 0,0469 | 0,9531 | 0,1406 |
| 270 | 0,47500 | 0,52500 | 0,3330 | 0,6670 | 0,1420 |
| 508 | 0,60000 | 0,40000 | 0,5333 | 0,4667 | 0,0667 |
| 746 | 0,75000 | 0,25000 | 0,6734 | 0,3266 | 0,0766 |
| 984 | 0,81250 | 0,18750 | 0,7714 | 0,2286 | 0,0411 |
| 1222 | 0,83750 | 0,16250 | 0,8401 | 0,1599 | 0,0026 |
| 1460 | 0,86250 | 0,13750 | 0,8881 | 0,1119 | 0,0256 |
| 1698 | 0,88750 | 0,11250 | 0,9217 | 0,0783 | 0,0342 |
| 1936 | 0,91250 | 0,08750 | 0,9452 | 0,0548 | 0,0327 |
| 2174 | 0,93750 | 0,06250 | 0,9616 | 0,0384 | 0,0241 |
| 2412 | 0,96250 | 0,03750 | 0,9732 | 0,0268 | 0,0107 |
| 2650 | 0,97500 | 0,02500 | 0,9812 | 0,0188 | 0,0062 |
| 2888 | 0,98750 | 0,01250 | 0,9869 | 0,0131 | 0,0006 |
| 3126 | 0,98750 | 0,01250 | 0,9908 | 0,0092 | 0,0033 |
| 3364 | 0,98750 | 0,01250 | 0,9936 | 0,0064 | 0,0061 |
| 3602 | 0,98750 | 0,01250 | 0,9955 | 0,0045 | 0,0080 |
| 3840 | 1,00000 | 0,00000 | 0,9968 | 0,0032 | 0,0032 |

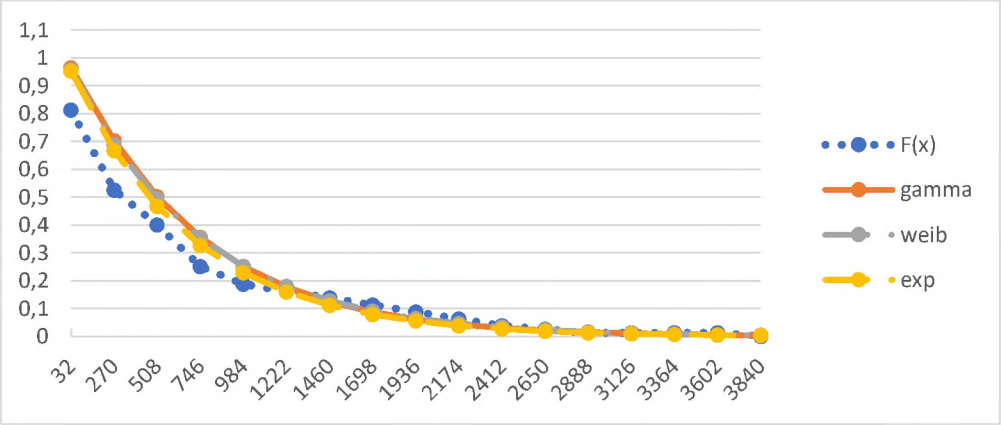



Рисунок 2.16 – Графіки емпіричної та відповідні їй теоретичні функції розподілу

Таким чином, результати перевірки адекватності моделі показали, що модель потоку задач можна апроксимувати експоненційним розподілом із

параметром mean = 712,0125. Але, якщо розглянути рисунку 2.16, то можна помітити, що починаючи із кількості задач 1200 на год. модель потоку задач також можна достатньо добре апроксимувати розподілами гамма і Вейбула із параметрами, що вказані у таблиці 2.3.

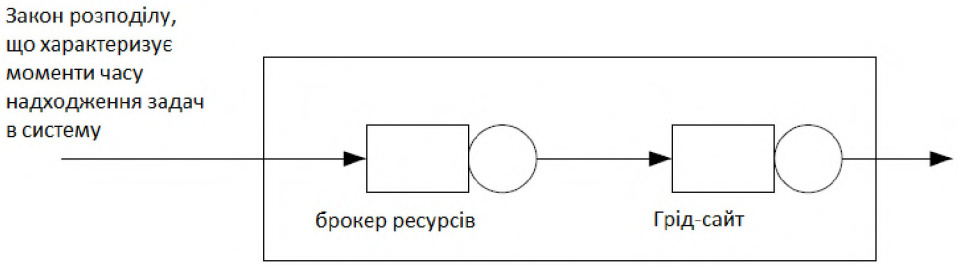



Рисунок 2.17 – Структурна схема ГРІД

Таким чином, отримані моделі інтенсивності потоку задач можуть використовуватися в якості вхідних параметрів моделювання ГРІД, що представлена на рисунку 2.17.

## 2.2 Розробка алгоритму

Нехай маємо дані про інтенсивність потоку задач у місяць (од./год)

Алгоритм побудови моделі потоку задач в ГРІД можна розбити на наступні кроки:

1. блок знаходження числових характеристик отриманої вибірки;

2. блок побудови ІВР, пошук функції щільності ймовірності;

3. блок побудови графіків;

4. блок пошуку НІ для математичного сподівання, дисперсії, середньоквадратичного відхилення з рівнем довіри *1-0,05*;

5. блок перевірити адекватність теоретичних розподілів (розподіли гамма, Вейбула, експоненціальний) за КК.

Опишемо докладніше деякі кроки. Блок пошуку числових характеристик включає пошук математичного сподівання, дисперсії, середньоквадратичного

відхилення, мінімального і максимального значень, коефіцієнта варіації, асиметрії, ексцесу. Блок побудови ІВР включає обчислення інтервалів, пошуку частот, відносних частот, функції щільності ймовірності. Блок побудови графіків включає побудову гістограми, графіків часового ряду, полігону відносних частот, функції щільності ймовірності, ccdf. Блок пошуку НІ. Довжина напівінтервалу для математичного сподівання обраховується за формулою

$$\Delta = \frac{t \cdot S}{\sqrt{n}}, \tag{2.39}$$

де $t$ шукаємо за таблицею Стьюдента з ступенями свободи який дорівнює $n-1$.

$$t = T_{кр}^{двост}(\alpha; n-1). \tag{2.40}$$

НІ дисперсії DX і середньоквадратичного відхилення σ шукаються за допомогою наступних формул:

$$\frac{n-1}{u_2} S^2 \le \sigma^2 \le \frac{n-1}{u_1} S^2, \tag{2.41}$$

$$\sqrt{\frac{n-1}{u_2} S^2} \le \sigma \le \sqrt{\frac{n-1}{u_1} S^2}, \tag{2.42}$$

де $u_1$, $u_2$ визначаються за таблицею критичних точок розподілу Пірсона для n-1-го степеня свободи:

$$u_1 = \chi^2_{кр}(1-\alpha/2, n-1);\ u_2 = \chi^2_{кр}(\alpha/2, n-1).$$

Блок перевірки адекватності теоретичного розподілу. Записуємо значення ccdf для емпіричного і теоретичних розподілів і знаходимо міру розходження розподілів за формулою (2.43)

$$D_n = \sup_{-\infty < x < \infty} \left| F_d(x) - F_e(x) \right| \tag{2.43}$$

і знаходимо величину

$$\lambda = D\sqrt{n} \tag{2.44}$$

Якщо величина (2.44) більша за критичне значення $\lambda_\alpha$, визначеного для рівня значущості $\alpha$, то нульова гіпотеза *H0* про те, що ВВ $X$ має заданий теоретичний розподіл відхиляється.

## 2.3 Побудова блок-схеми алгоритму

На рисунку 2.18 показаний алгоритм побудови моделі потоку задач у ГРІД.

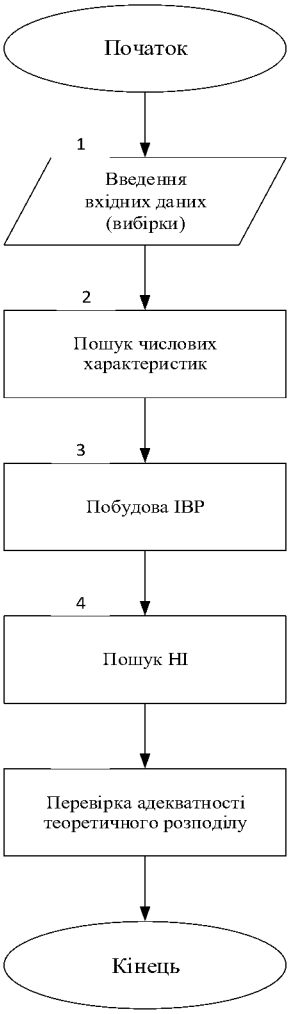



Рисунок 2.18 – Обчислювальний алгоритм побудови моделі потоку задач в ГРІД

На рисунку 2.19 показана схема перевірки адекватності теоретичного розподілу за допомогою КК.

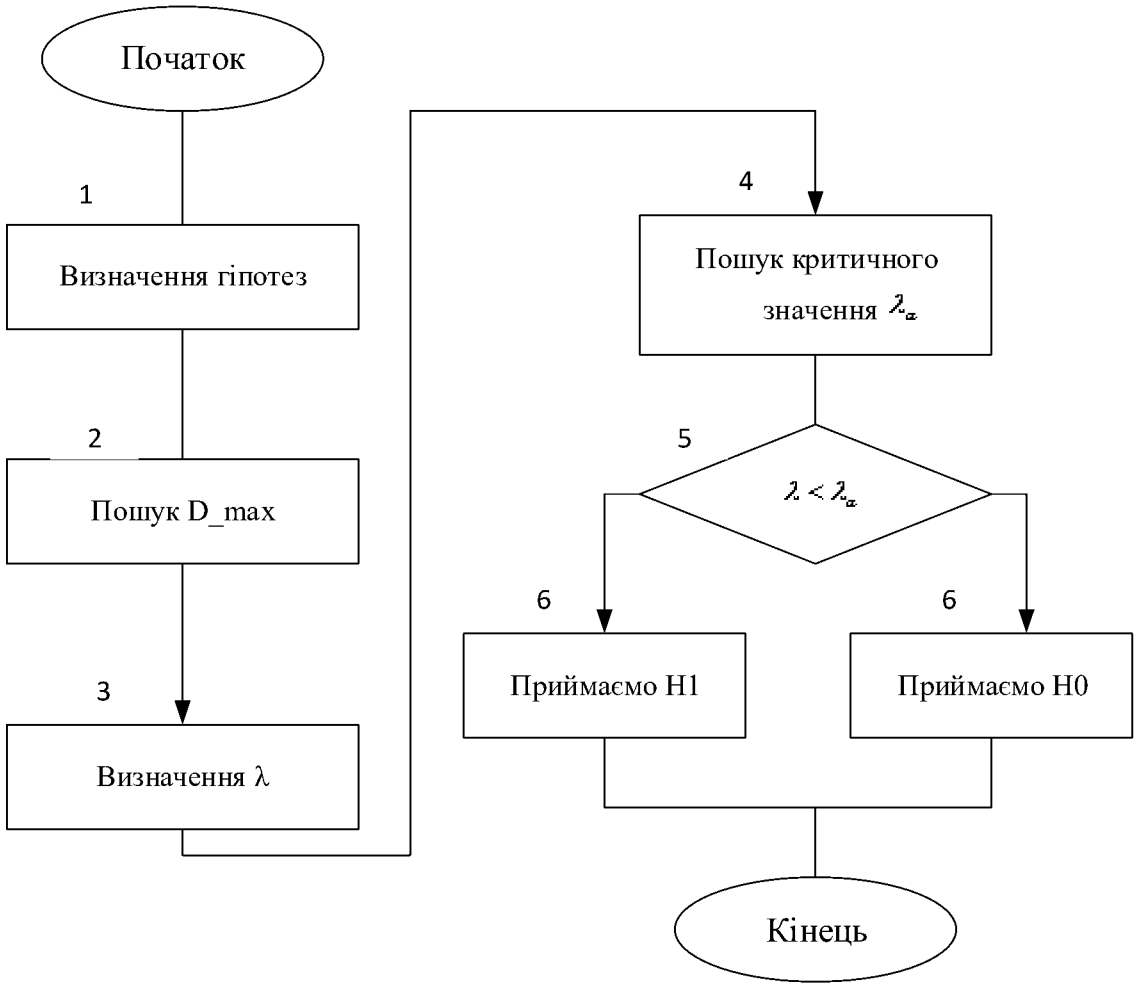



Рисунок 2.19 – Алгоритм перевірки теоретичного розподілу за критерієм Колмогорова

## 2.4 Інструментальні засоби і технології

Для розробки програмної реалізації побудови моделі потоку задач в ГРІД я використав мову C#.

C# (вимовляється Сі-шарп) – об'єктно-орієнтована мова програмування з безпечною системою типізації для платформи .NET. Розроблена Андерсом Гейлсбергом, Скотом Вілтамутом та Пітером Гольде під егідою Microsoft Research (при фірмі Microsoft).

Синтаксис C# близький до C++ і Java. Мова має строгу статичну типізацію, підтримує поліморфізм, перевантаження операторів, вказівники на функції-члени класів, атрибути, події, властивості, винятки, коментарі у форматі XML.

Перейнявши багато що від своїх попередників — мов C++, Delphi, Модула і Smalltalk — C#, спираючись на практику їхнього використання, виключає деякі моделі, що зарекомендували себе як проблематичні при розробці програмних систем, наприклад множинне спадкування класів (на відміну від C++).

Для реалізації поставленої задачі я використав середовище програмування Visual Studio 2010 – продукт фірми Майкрософт, який включає інтегроване середовище розробки програмного забезпечення та ряд інших інструментальних засобів. Цей продукт дозволяють розробляти як консольні програми, так і програми з графічним інтерфейсом, в тому числі з підтримкою технології Windows Forms, а також веб-сайти, веб-додатки, веб-служби як в рідному, так і в керованому кодах для всіх платформ, що підтримуються Microsoft Windows, Windows Mobile, Windows CE, .NET Framework, .NET Compact Framework та Microsoft Silverlight. У Visual Studio можна розробляти програми на таких мовах програмування: C++, C#, Visual Basic, F#.

# РОЗДІЛ 3
# РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

## 3.1 Програмна реалізація

Для розробки програми було використано мову програмування C#, тип проєкта - Windows Forms.

Для вирішення поставленої задачі я реалізував наступні класи:

- NumericalCharacteristics – клас, в якому реалізовані функції для пошуку числових характеристик;

- IntervalVariation – клас, в якому реалізовані функції для побудови IBP;

- CredibleIntervals – клас, в якому реалізовані функції для пошуку HI;

- MainForm – клас, який описує головне вікно програми;

- KolmogorovGraph – клас, який описує вікно, в якому будуються графіки ccdf для ЕФР і ТФР.

Розглянемо класи детальніше.

Клас NumericalCharacteristics має наступні функці:

- public int Max() – обчислює максимальне значення у вибірці;

- public int Min() – обчислює мінімальне значення у вибірці;

- public int Count() – обчислює кількість елементів у вибірці;

- public double ExpectedValue() – обчислює математичне сподівання;

- public double Variance() – обчислює дисперсію;

- public double CorrectedVariance() – обчислює виправлену дисперсію;

- public double StandardDeviation() – обчислює середньоквадратичне відхилення;

- public double CorrectedDeviation() – обчислює виправлене середньоквадратичне відхилення;

- public double CoefficientOfVariation() – обчислює коефіцієнт варіації;

- public double Skewness() – обчислює коефіцієнт асиметрії;

- public double Kurtosis() – обчислює коефіцієнт ексцесу.

Розглянемо основні функції у класі IntervalVariation:

- public List<int> GetXi() – повертає варіанти;

- public List<int> GetAi() – повертає список початків інтервалів;

- public List<int> GetBi() – повертає список кінців інтервалів;

- public  List<int> GetNi() – повертає список частот;

- public List<double> GetFrequency() - повертає список відносних частот;

- public List<int> GetNi_0() – повертає список накопичених частот.

- public List<double> GetRelativeFrequency() – повертає список накопичених відносних частот;

- public List<double> GetValueH() – повертає список висот прямокутників для побудови гістограми;

- public List<double> GetCCDF() – повертає значення ccdf;

- private int CalculateValueK() – обчислює довжину інтервалів.

У класі CredibleIntervals реалізовані наступні функції:

- public void CalculateIntervalsForExpectation(out double x1, out double x2) – повертає HI для математичного сподівання;

- public void CalculateIntervalsForVariance(out double x1, out double x2) – повертає HI для дисперсії;

- public void CalculateIntervalsForStandardDeviation(out double x1, out double x2) – повертає HI для середньоквадратичного відхилення.

Клас KolmogorovGraph призначений для побудови порівняльних графіків при перевірці теоретичного розподілу для КК – для цього у класі був реалізований метод public KolmogorovGraph(List<double> xi, List<double> ccdfFdX, List<double> ccdfFeX). Функція InitializeComponent() ініціалізує компоненти на даній формі.

Клас MainForm описує головне вікно програми. Для зручності зображення отриманих даних на формі я використав компонент TabControl.Опишемо основні функції, що були реалізовані в даному класі:

- private void InitSampleDataGridView(List<int> numbers) – виведення на форму таблиці із вибіркою;

- private void InitNumCharacteristicsDataGridView(List<int> numbers) – виведення на форму таблиці, в якій перечислені знайдені числові характеристики;

- private void DrawGraph(List<double> xi, List<double> yi, string name, bool isHistogram) – побудова графіків;

- private void InitCredibleIntervals(List<int> numbers – виведення на форму HI.

### 3.2 Використання програмного продукту

Розроблена програма досить легка у користуванні. Вхідні дані для програми потрібно записати у текстовий файл. Для зчитування даних потрібно у меню «Файл» клікнути на «Відкрити файл». Після відкриття файлу головне вікно програми буде мати вид, показаний на рисунку 3.20. Вкладка «Числові характеристики» містить числові характеристики для заданої вибірки. У вкладці «ІВР» містяться дані при побудові ІВР, зокрема варіанти, інтервали, частоти, накопичені частоти, відносні частоти, відносні накопичені частоти, значення ccdf (рис. 3.21).

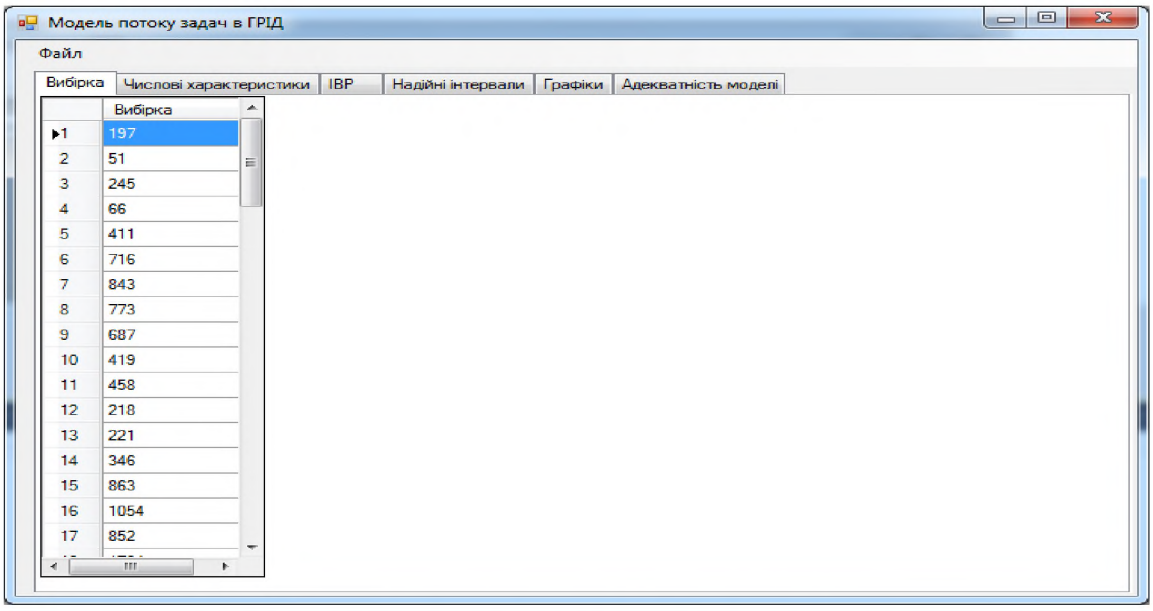
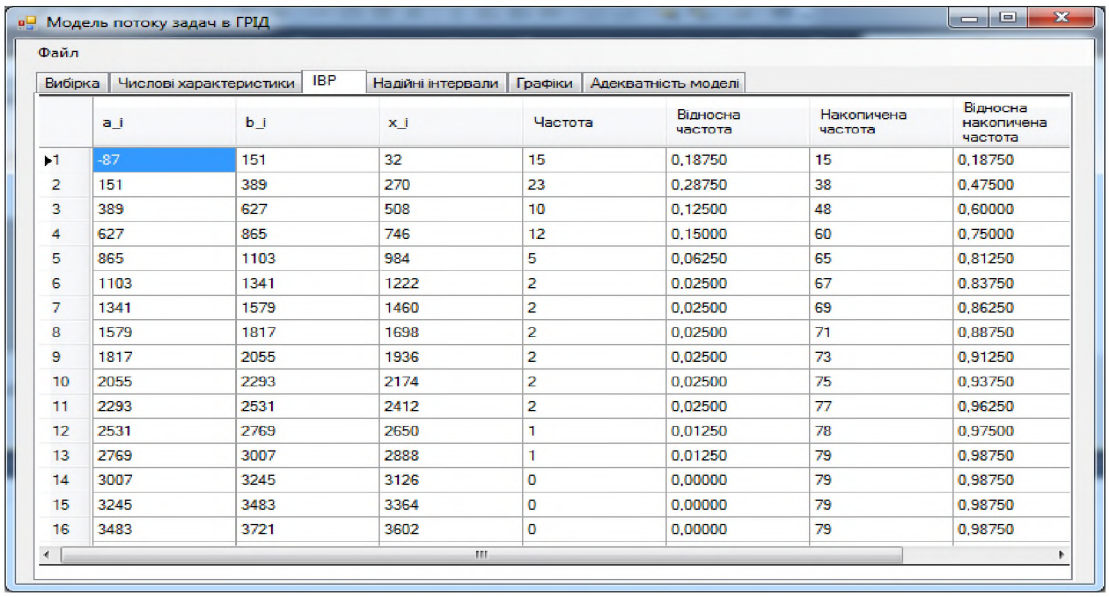



Рисунок 3.20 – Головне вікно програми




| | a_i | b_i | x_i | Частота | Відносна частота | Накопичена частота | Відносна накопичена частота |
|---|---|---|---|---|---|---|---|
| ▸1 | -87 | 151 | 32 | 15 | 0,18750 | 15 | 0,18750 |
| 2 | 151 | 389 | 270 | 23 | 0,28750 | 38 | 0,47500 |
| 3 | 389 | 627 | 508 | 10 | 0,12500 | 48 | 0,60000 |
| 4 | 627 | 865 | 746 | 12 | 0,15000 | 60 | 0,75000 |
| 5 | 865 | 1103 | 984 | 5 | 0,06250 | 65 | 0,81250 |
| 6 | 1103 | 1341 | 1222 | 2 | 0,02500 | 67 | 0,83750 |
| 7 | 1341 | 1579 | 1460 | 2 | 0,02500 | 69 | 0,86250 |
| 8 | 1579 | 1817 | 1698 | 2 | 0,02500 | 71 | 0,88750 |
| 9 | 1817 | 2055 | 1936 | 2 | 0,02500 | 73 | 0,91250 |
| 10 | 2055 | 2293 | 2174 | 2 | 0,02500 | 75 | 0,93750 |
| 11 | 2293 | 2531 | 2412 | 2 | 0,02500 | 77 | 0,96250 |
| 12 | 2531 | 2769 | 2650 | 1 | 0,01250 | 78 | 0,97500 |
| 13 | 2769 | 3007 | 2888 | 1 | 0,01250 | 79 | 0,98750 |
| 14 | 3007 | 3245 | 3126 | 0 | 0,00000 | 79 | 0,98750 |
| 15 | 3245 | 3483 | 3364 | 0 | 0,00000 | 79 | 0,98750 |
| 16 | 3483 | 3721 | 3602 | 0 | 0,00000 | 79 | 0,98750 |

Рисунок 3.21 – Вкладка IBP

Дані про НІ знаходяться у відповідній вкладці.

У вкладці «Графіки» можна розглянути графіки для обраної вибірки. Для цього потрібно вибрати графік у випадаючому списку і побудувати його (рис. 3.22).

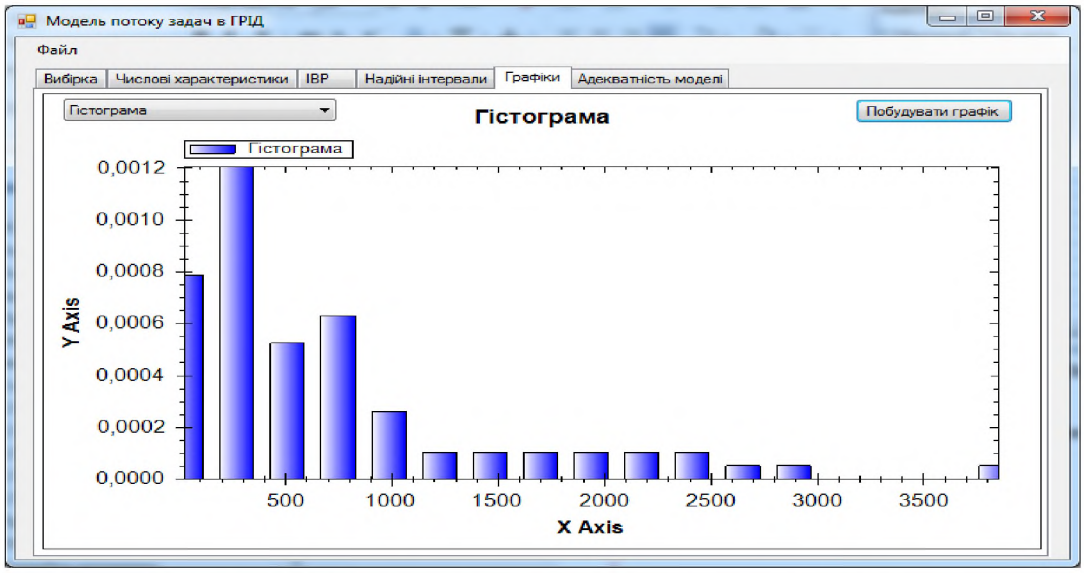



Рис. 3.22 – Графік гістограми для заданої вибірки

Остання вкладка служить для перевірки адекватності вибраної моделі. У випадаючому списку вибираємо теоретичний розподіл, який будемо оцінювати

на відповідність із заданим емпіричним розподілом (значення відносної накопиченої частоти). На рисунку 3.23 представлена перевірка адекватності моделі для розподілу Вейбула.

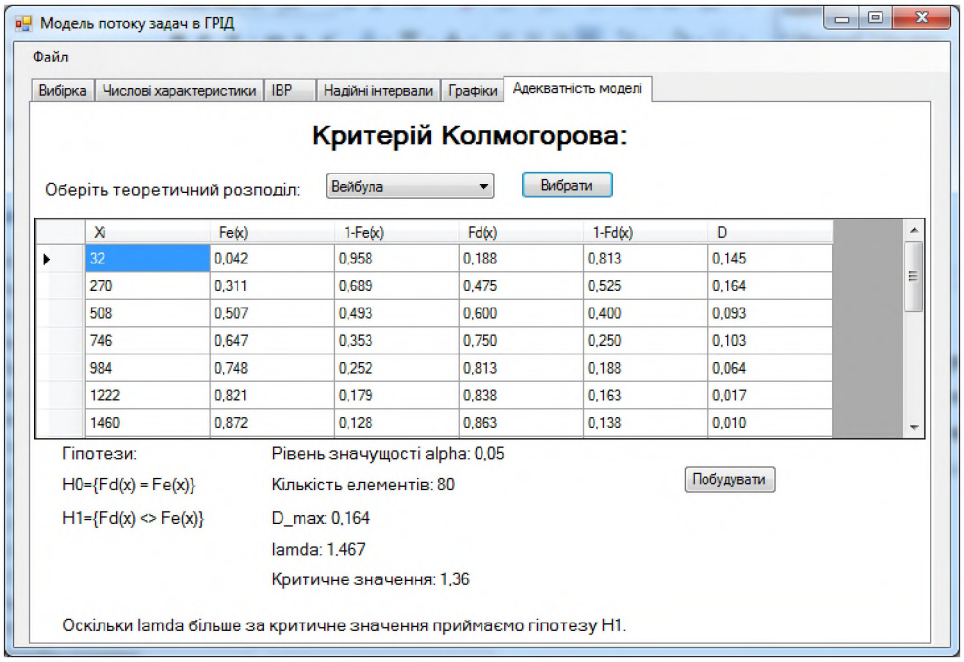



Рисунок 3.23 – Перевірка адекватності теоретичного розподілу

## 3.3 Вартість розробки

Вартість розробки розрахована в таблиці 3.7. Вартість часу, витраченого на розробку, взяті з середніх показників погодинної вартості роботи на тематичних вебресурсах.

Таблиця 3.7 – Вартість розробки

| Складова розробки | Вартість | Пояснення |
|---|---|---|
| Вартість програмного забезпечення, використаного для розробки | 0 грн. | Програмне забезпечення, яке використовувалося при розробці – безкоштовне |
| Вартість часу, витраченого на розробку | 20 год х 600 грн. | Усереднене значення рівня розробника нижче середнього на тематичних сайтах |

# ВИСНОВКИ

На основі проведеної роботи, можна зробити наступні висновки.

1. Розглянуто архітектуру ГРІД.

2. Розглянуто статистичні методи побудови моделі потоку задач в ГРІД в теорії і практиці.

3. Розроблено програмний продукт для побудови моделі потоку задач в ГРІД.

Таким чином, одним із механізмів збільшення відмовостійкості у ГРІД системах є розуміння характеру потоку задач і його розподілу між ресурсами на основі деякого критерію оптимізації (наприклад, мінімізації часу виконання задачі). Так, як значення параметрів робочого навантаження (потоку задач) у ГРІД носить ймовірнісний характер, то доцільно проводити їх дослідження за допомогою методів математичної статистики, що дозволяють чисельно і графічно описати дані, а також побудувати модель потоку задач.

ГРІД-обчислення – це форма розподілених обчислень, в якій «віртуальний суперкомп'ютер» представлений у виді кластера, з'єднаних за допомогою мережі, комп'ютерів, що працюють разом для виконання наукових, математичних задач, що потребують значних обчислювальних ресурсів. За допомогою ГРІД виконуються деякі трудомісткі завдання, пов'язані з економічною прогнозуванням, сейсмоаналізом, розробкою та вивченням властивостей нових ліків.