Лекція 10

Первинний статистичний аналіз

1. Застосування статистики при аналізі результатів вимірювань ПЗ.
2. Первинний статистичний аналіз.
3. Закон розподілу.
4. Статистичні перевірки.

Проблема аналізу вимірювань

- На основі вимірювання простих властивостей програмного забезпечення потрібно робити висновки про загальні його властивості

Застосування статистичного аналізу для ПЗ

- Ідентифікація розподілу
- Пошук та відображення залежностей між даними
- прогнозування

Вибірка

- Це деякий набір значень величини із загальної кількості її значень (генеральної сукупності).

- Достатність вибірки – представлення вибіркою генеральної сукупності (при збільшенні об’єму даних середні статистичні характеристики змінюються несуттєво)

Гістограми

- Побудова варіаційного ряду (гістограми) вимагає ранжування результатів спостережень та обчислення відповідних їм частот і випадковостей:

$$x_1, \quad x_2, \quad \ldots, \quad x_r$$
$$n_1, \quad n_2, \quad \ldots, \quad n_r$$
$$f_1, \quad f_2, \quad \ldots, \quad f_r,$$

- де r – кількість варіант;
- x_i – і-те значення х метрики;
- n_i – частота x_i, ;
- - випадковість x_i.

Гістограми

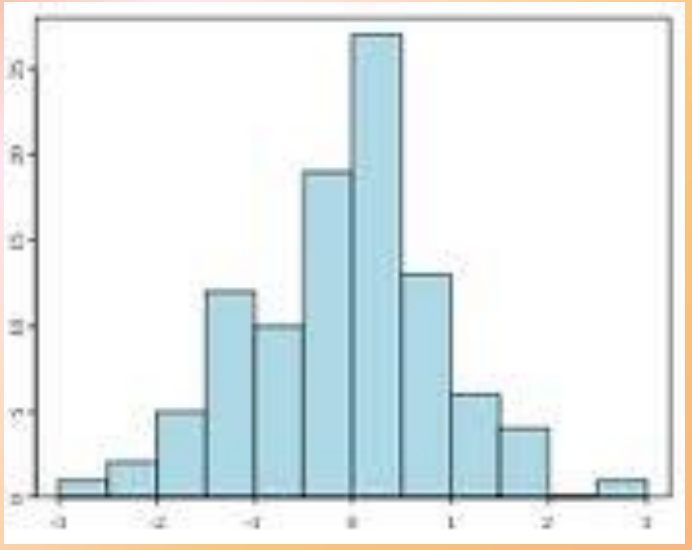
- Для побудови гістограми проводиться розбиття варіаційного ряду на класи. Для цього фіксується рівномірне розбиття осі спостережень Δ_h на класи, де h — крок розбиття. Крок розбиття визначається із співвідношення:
- $$h = \frac{b-a}{m}$$
- a — початок спостережень (окремий випадок x_1 = a);
- b — кінець спостережень (окремий випадок $x_г$ = b);
- *m* — кількість елементів розбиття Δ_h (кількість

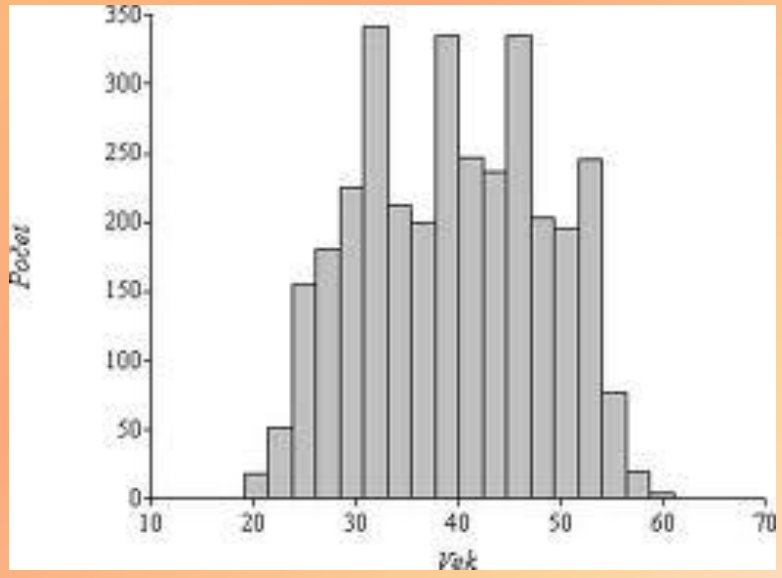
Гістограми

- Кількість класів — величина довільна.
- Краще вибирати *m* непарним і таким, щоб гістограма, по можливості, не мала осциляції випадковостей і була більш-менш "гладкою".
- Існує оптимальна кількість класів, яка залежить від обсягу даних вибірки *n* та від типу їх закону розподілу (мається на увазі врахування асиметрії та ексцесу). При n < 100 можна використати формулу

$$m = \begin{cases} [\sqrt{n}], & \text{коли } [\sqrt{n}] - \text{парне}, \\ [\sqrt{n}] - 1, & \text{коли } [\sqrt{n}] - \text{непарне}, \end{cases}$$

Гістограми





Аналіз неперервних та дискретних даних

- Неперервні дані представляються у вигляді функцій
- При аналізі дискретні дані краще представляти у неперервній формі

Математичне сподівання

- Середнє арифметичне, яке є оцінкою математичного сподівання випадкової величини

$$\bar{x} = \hat{\nu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{r} x_i n_i = \sum_{i=1}^{r} x_i f_i$$

Дисперсія та середнє квадратичне відхилення

- Вибіркова дисперсія та середньоквадратичне відхилення характеризує розсіювання вибіркових даних відносно середнього

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{r}(x_i - \bar{x})^2 n_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 f_i, \sigma = S$$

Коефіцієнти асиметрії та ексцесу

- Коефіцієнт асиметрії, що характеризує асиметричність функції щільності (гістограми) відносно середнього

$$\hat{A} = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{1}{n\hat{\sigma}^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{n\hat{\sigma}^3}\sum_{i=1}^{r}(x_i - \bar{x})^3 n_i = \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i$$

- Коефіцієнт ексцесу характеризує гостровершинність функції розподілу (гістограми) відносно нормального

$$\hat{E} = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{1}{n\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 = \frac{1}{n\hat{\sigma}^4}\sum_{i=1}^{r}(x_i - \bar{x})^4 n_i = \frac{1}{\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 f_i$$

Довірчі інтервали

- Використовується для оцінювання точності оцінок параметрів

$$\hat{\theta} - t_{\alpha/2,\nu}\sigma\{\hat{\theta}\} < \theta < \hat{\theta} + t_{\alpha/2,\nu}\sigma\{\hat{\theta}\},$$

- $t_{\alpha/2,\nu}$ – квантиль t-розподілу Стьюдента.
- За величину беруть відповіді точкову оцінку, а значення а визначають із співвідношень:

$$\sigma\{\bar{x}\} = \frac{S}{\sqrt{n}},$$

$$\sigma\{S\} = \frac{S}{\sqrt{2n}},$$

$$\sigma\{\bar{A}\} = \sqrt{\frac{6}{n}\left(1 - \frac{12}{2n+7}\right)} \text{ або } \sigma\{\bar{A}\} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}},$$

$$\sigma\{\bar{E}\} = \sqrt{\frac{24}{n}\left(1 - \frac{225}{15n+124}\right)} \text{ або } \sigma\{\bar{E}\} = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}},$$

Вилучення аномальних значень

- Обчислені значення статистики

$$t = \frac{|x_{гр} - \bar{x}|}{S}$$

- Порівнюється з критичним значенням $t_{a/2,v}$ (квантиль розподілу Стьюдента)
- При $t \geq t_{a/2,v}$, ...ідлягає видаленню

Вилучення аномальних значень

- Підсумком аналізу варіаційного ряду або гістограми може бути попередній висновок про наявність аномальних ("грубих") значень $x_{гр}$.
- Візуально такі значення можна ідентифікувати з аналізу гістограм, коли значення варіаційного ряду досить суттєво віднесене від загальної сукупності даних та має порівняно малу випадковість.
- Варіанта x_i за своїм значенням може різко відхилятися від загальні сукупності варіант у двох випадках:
  - якщо вона належить до генеральної сукупності, як і основна група, проте є малоймовірною подією
  - або якщо має місце випадкове порушення умов експерименту.

Види розподілів

- Однопараметричні
  - Експоненційний
  - Релея
  - Максвела
  - Пірсона
  - Т-розподіл Стьюдента
- Двопараметричні
  - Рівномірний
  - Паретто
  - Нормальний
  - Логарифмічно-нормальний
  - Лапласа
  - Гамма-розподіл
  - Екстремальний
  - Розподіл Вейбула

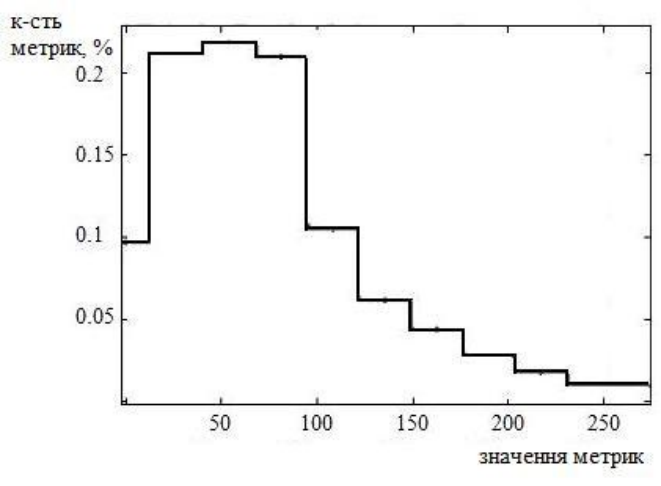
Закон розподілу

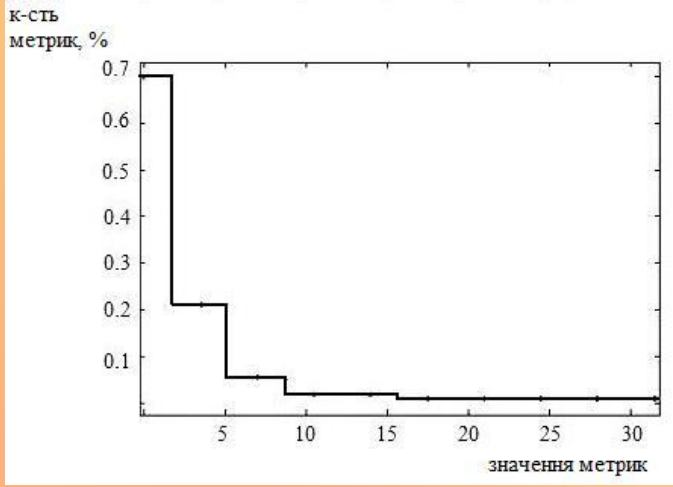
- Використовується для дискретної випадкової величини
- Показує множину можливих подій з ймовірностями їх настання

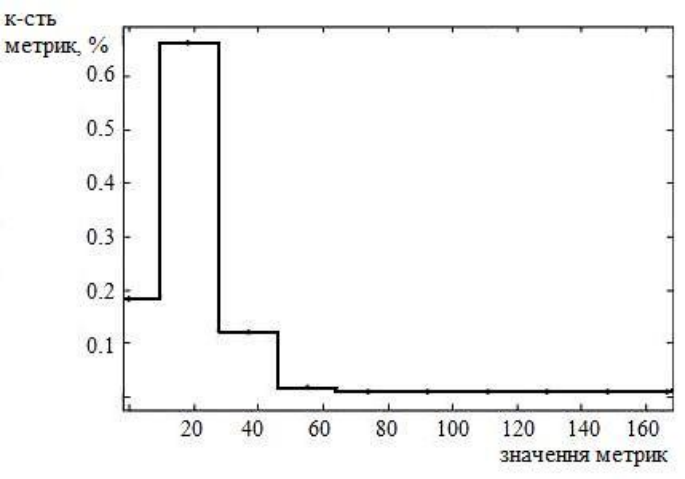
Ідентифікація розподілів (крок 1)

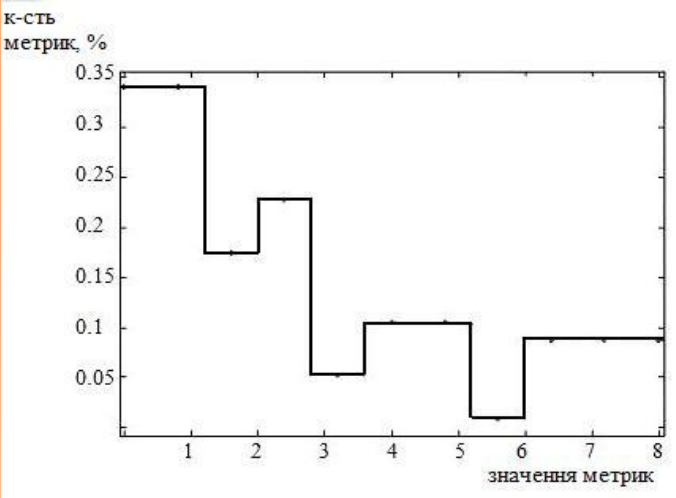
- На практиці при первинному статистичному аналізі тип розподілу невідомий
- Попередньо проводять ідентифікацію, аналізуючи гістограму (крок 1)

Ідентифікація розподілів









Ідентифікація розподілів

- Унімодальна гістограма:
  - Експоненційний
  - Вейбула з параметром $\beta <= 1$
  - Паретто
  - …
- Симетрична гістограма:
  - Нормальний
  - Розподіл Стьюдента
  - Лапласа
  - Коші
  - Релея
- Одномодальна асиметрична гістограма:
  - Логарифмічно-нормальний
  - Вейбула з параметром $\beta > 1$

Ідентифікація розподілів (крок 2)

- Вибір конкретного типу розподілу за емпіричною функцією розподілу (крок 2)
- 2 підходи:
  - Перетворення функції розподілу для надання лінійного вигляду (переважно – перетворення Джонсона)
  - Моментна ідентифікація – за допомогою коефіцієнтів асиметрії та ексцесу

Ідентифікація розподілів – моментні характеристики

Розподіл	A	E
Нормальний	0	0
Експоненційний	2	6
Максвелла	0,065375	1,569972
Рівномірний	0	1,2
Лапласа	2,12132	3
Екстремальний	1,12396	2,4

Вибір розподілу базується на перевірці гіпотези відхилення емпіричних значень від заданих в таблиці

Уточнення розподілу здійснюється на основі критеріїв згоди

Відтворення розподілів

- Метою відтворення розподілів є побудова функції розподілу за вибірковими даними

Схема відтворення розподілів

Основні кроки

- 1. Первинний статистичний аналіз
- 2. Знаходження оцінок параметрів
- 3. Оцінювання точності оцінок параметрів шляхом обчислення дисперсії та довірчих інтервалів
- 4. Обчислення значень статистичної функції розподілу у точках варіаційного ряду
- 5. Визначення одного або кількох критеріїв згоди
- 6. Довірче оцінювання теоретичної функції розподілу ймовірностей

Схема відтворення розподілів

Первинний статистичний аналіз

- Формування варіаційних рядів
- Розбиття варіаційних рядів на класи
- Вилучення аномальних значень
- Обчислення емпіричної функції розподілу ймовірностей
- Знаходження статистичних характеристик вибірки з довірчим оцінюванням
- Ідентифікація типу розподілу

Методи оцінки параметрів розподілу

- Метод максимальної правдоподібності – відбувається порівняння емпіричних та теоретичних статистичних характеристик
- Метод моментів – базується на порівнянні теоретичних та статистичних початкових або центральних моментів
- Метод найменших квадратів – використовується при ефективному перетворенні функції розподілу до

Висновки

- Статистичний аналіз найбільш використовується при аналізі деяких вибірок даних