----------МАТЕРІАЛ З ПОСІБНИКУ НА 24 СТОРІНЦІ ФАЙЛУ----------

# Попередня обробка даних

***Попередня обробка даних*** займається знаходженням по спостереженням змінних їх первинних характеристик та допоміжної інформації, які стануть у нагоді у подальших розділах обробки та аналізу інформації (даних).

До попередньої обробки даних входять:

- *підрахування базових характеристик розподілу змінної*, яка спостерігається (а саме: початкових та центральних моментів, квантилів та відсоткових точок, характеристик центру значень змінної, характеристик розсіювання значень змінної, асиметрії, ексцесу, емпіричної функції розподілу, емпіричної функції щільності і т.п.),
- *виявлення та вилучення аномальних спостережень*,
- *перевірка основних гіпотез* (а саме: стохастичності вибірки, симетрії розподілу, згоди з заданим законом розподілу, однорідності вибірки, і т.п.),
- *розвідувальний аналіз* (проводить візуальний попередній експрес-аналіз інформації).

## Базові характеристики

До них відносять: початкові та центральні моменти, квантилі та відсоткові точки, характеристики центру значень змінної, характеристики розсіювання значень змінної, асиметрію, ексцес, емпіричну функцію розподілу, емпіричну функцію щільності та інші характеристики, які будуть служити хорошим підґрунтям для подальшого аналізу даних.

Згадаємо деякі поняття.

**Означення**. ***Варіаційним рядом*** вибірки

$$x_1, x_2, \ldots, x_n$$

називається така послідовність

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)},$$

яка утворена з цієї вибірки після розташування її значень у порядку не спадання, тобто справедливо

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}.$$

**Означення**. Члени варіаційного ряду

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

називаються ***порядковими статистиками***.

**Означення**. $i$ – ий член варіаційного ряду $x_{(i)}$ називається $i$ – ***ою порядковою статистикою***.

## Квантилі та відсоткові точки розподілу

Ці поняттями будуть широко використовуватися у подальшому при розв'язанні задач перевірки гіпотез, побудові довірчих інтервалів та областей і т.п. Визначати їх будемо окремо для неперервних та дискретних розподілів. Спочатку дамо означення для теоретичних значень квантилів.

**Означення**. ***[Теоретичним] квантилем рівня*** $q$ ***розподілу*** ($q$-квантилем розподілу) ***неперервної випадкової величини*** $\xi$ називається таке дійсне число $u_q$, яке визначається з рівняння

$$P\{\xi < u_q\} = q, \quad 0 < q < 1.$$

**Означення**. ***[Теоретичним] квантилем рівня*** $q$ ***розподілу*** ($q$-квантилем розподілу) ***неперервної випадкової величини*** $\xi$ називається таке дійсне число $u_q$, яке визначається з рівняння

$$P\{\xi < u_q\} = q, \quad 0 < q < 1.$$

**Означення**. ***[Теоретичним] квантилем рівня*** $q$ ***розподілу*** ($q$-квантилем розподілу) ***дискретної випадкової величини*** $\xi$ з варіаційним рядом своїх значень $\{y_{(i)}\}$ називається довільне значення $u_q$ з інтервалу $\left(y_{(i(q))}, y_{(i(q)+1)}\right]$, для границь якого справедливо

$$P\{\xi < y_{(i(q))}\} < q,$$

$$P\{\xi < y_{(i(q)+1)}\} \ge q, \quad (0 < q < 1).$$

**Означення**. ***Емпіричним (вибірковим) квантилем рівня*** $q$ ***розподілу*** випадкової величини $\xi$ називається квантиль рівня $q$ відповідного емпіричного (вибіркового) розподілу.

**Приклади квантилів**.

1.***Медіана*** – це квантиль рівня 0,5, тобто $u_{0,5}$.

2.***Нижній та верхній квартилі*** визначаються як $u_{0,25}$ та $u_{0,75}$ відповідно.

3.***Децилі*** – це квантилі $\left\{u_{\frac{i}{10}}\right\}_{i=1}^{9}$.

4.***Центилі*** задаються наступним чином $\left\{u_{\frac{i}{100}}\right\}_{i=1}^{99}$.

5.***Інтерквантильна широта рівня*** $q$ $\left(0<q<\frac{1}{2}\right)$ – це величина, яка обчислюється по формулі

$$\left(u_{1-q}-u_q\right).$$

6.***Інтерквартильна широта*** – це інтерквантильна широта рівня 0.25, а саме

$$\left(u_{0,75}-u_{0,25}\right).$$

7.***Ймовірнісне відхилення*** $d_\xi$ визначається як половина інтерквартильної широти, тобто $d_\xi=\frac{1}{2}\left(u_{0,75}-u_{0,25}\right)$.

8.***Інтерсектильна широта*** – це інтерквантильна широта рівня $\frac{1}{6}$, тобто

$$\left(u_{\frac{5}{6}}-u_{\frac{1}{6}}\right).$$

9.***Інтердецильна широта*** – це інтерквантильна широта рівня 0.1, а саме $\left(u_{0,9}-u_{0,1}\right)$.

**Означення**. ***[Теоретичною] $Q$-відсотковою точкою розподілу неперервної випадкової величини*** $\xi$ називається таке дійсне число $v_Q$, яке є розв'язком рівняння

$$P\{\xi \geq v_Q\} = \frac{Q}{100}, \quad 0 < Q < 100.$$

**Означення**. ***[Теоретичною] Q-відсотковою точкою розподілу дискретної випадкової величини*** $\xi$ з варіаційним рядом своїх значень $\{y_{(i)}\}$ називається довільне значення $v_Q$ з інтервалу $\left(y_{(i(Q))}, y_{(i(Q)+1)}\right]$, для границь якого справедливо

$$P\{\xi \geq y_{(i(Q))}\} > \frac{Q}{100},$$

$$P\{\xi \geq y_{(i(Q)+1)}\} \leq \frac{Q}{100}, \quad 0 < Q < 100.$$

**Означення**. ***Емпіричною (вибірковою) $Q$-відсотковою точкою розподілу*** випадкової величини $\xi$ називається Q-відсоткова точка відповідного емпіричного (вибіркового) розподілу.

Ці два поняття взаємно доповнюють одне одного. У неперервному випадку для певного розподілу взаємозв'язок між ними прозорий і має наступний вигляд:

$$u_q = v_{(1-q)100}, \quad v_Q = u_{1-\frac{Q}{100}}.$$

Для широко вживаних розподілів складені відповідні таблиці, з яких легко визначити потрібні квантилі та відсоткові точки.

**Самостійна робота №2**. З навчального посібника «Слабоспицький О.С. Аналіз даних. Попередня обробка, 2001» необхідно пропрацювати матеріал наведений у **розділах 2.1-2.5**:

Характеристики положення центра значень змінної.

Характеристики розсіювання значень змінної.

Аналіз скошеності та гостроверхості розподілу.

Характеристики випадкових векторів.

# Виявлення та вилучення аномальних спостережень

*Причини появи у вибірках аномальних спостережень (викидів):* збій у роботі обладнання, суттєві похибки вимірювальних приладів, порушення умов проведення експерименту, стихійні лиха, форс-мажорні обставини, інші непередбачувані причини і т.п.

**Означення**. ***Аномальними спостереженнями (викидами)*** у вибірці називаються ті виміри з неї, значення яких не узгоджуються з розподілом більшості отриманих спостережень.

Виявлені у вибірці аномальні спостереження, при наявності можливості *корегують*, інакше, як правило, їх *видаляють* з вибірки, *але обережно*. Бо головне при цьому «не вихлюпнути разом з водою і дитя», бо саме екстремальні виміри, які у більшості випадків і підозрюють на аномальність, дуже часто є найбільш інформативними спостереженнями. Саме спостереження отримані *під час функціонування досліджуємого об'єкту в екстремальних режимах*, як правило, і несуть найбільше інформації про цей об'єкт.

**Приклад 2.** Випробування літальних апаратів. А саме …

Так як найбільш розробленою виявилася теорія *для нормальних вибірок*, то саме її і будемо розглядати у подальшому.

**I. *Випадок обробки скалярних спостережень.*** Нехай маємо

$$\xi: x_1, x_2, \ldots, x_n.$$

Потрібно виявити та вилучити викиди з цієї вибірки. Для цього було запропоновано ряд методів, а саме:

1) критерій Граббса,
2) критерій Томпсона,
3) критерій Тітьєна-Мура,
4) засоби розвідувального аналізу:
   - пробіт-графік,
   - ймовірнісний графік,
   - зображення "скринька з вусами",
   - зображення "стебло-листок".

Зупинимося на кожному з них більш детальніше.

**Зауваження**. У подальшому при перевірці деякої гіпотези $H_0$, якщо не буде згадуватися протилежна гіпотеза $H_1$, то по замовчуванню будемо вважати, що вона є простим запереченням $H_0$.

**1) Критерій Граббса.** Тут задачу розв'язують шляхом перевірки:
$H_0$ : *найбільш підозрілий на аномальність вимір не є викидом*, $\alpha > 0$.

1. Спочатку за вибіркою підрахуємо такі статистики:

$$\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s(n) = \sqrt{\frac{1}{n}\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2(n)\right]}.$$

2. Далі будуємо послідовність:

$$z_1, z_2, \quad \ldots, \quad z_n,$$

де $z_i = \left|x_i - \bar{x}(n)\right|, \ i = \overline{1,n}.$

3. А потім відповідний варіаційний ряд:

$$z_{(1)}, z_{(2)}, \quad \ldots, \quad z_{(n)}.$$

де $z_{(j)} = \left|x_{i(j)} - \bar{x}(n)\right|, \ j = \overline{1,n}.$

---

Підозрюємо на аномальність спостереження, яке відповідає останньому члену варіаційного ряду $z_{(n)}$, а саме $x_{i(n)}$. Перевіримо його на аномальність.

4. Для цього обчислимо таку статистику:

$$T(n) = \frac{x_{i(n)} - \bar{x}(n)}{s(n)}.$$

5. Тоді логічно за *область прийняття гіпотези* $H_0$, що $x_{i(n)}$ не є викидом, взяти область, яка не включає у себе екстремальних значень останньої статистики, а саме:

$$\left|T(n)\right| < T_{\frac{\alpha}{2}}(n),$$

де $T_{\frac{\alpha}{2}}(n)$ - $100\dfrac{\alpha}{2}$ відсоткова точка розподілу статистики $\dfrac{x_{i(n)} - \bar{x}(n)}{s(n)}$.

Якщо $x_{i(n)}$ є аномальним, то його видаляють з вибірки і всю процедуру повторюють починаючи з пункту 1, але вже з отриманою

скороченою вибіркою. Все це повторюється до тих пір, доки на деякому кроці найбільш підозрілий на аномальність вимір виявиться не викидом. Після цього роботу завершують.

**Зауваження 1.** Відповідна таблиця цих відсоткових точок отримана Граббсом і наведена у довіднику:
Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.

**Зауваження 2.** Критерій Граббса, в основному, використовується для невеликих значень $n$.

**2) Критерій Томпсона.** Він є модифікацією критерію Грабса. Знову задача розв'язується шляхом перевірки гіпотези:
$H_0$ : *найбільш підозрілий на аномальність вимір не є викидом*, $\alpha > 0$.

1. ……
2. ……
3. ……
4. Для цього обчислимо таку статистику:

$$T(n) = \frac{x_{i(n)} - \bar{x}(n)}{s(n)}.$$

5. А далі будується статистика:

$$t(n) = \frac{\sqrt{n-2}\,T(n)}{\sqrt{n-1-T^2(n)}},$$

розподіл якої можна наблизити $t$ – розподілом Стьюдента з $(n-2)$ степенями свободи при достатньо великих $n$.

6. Аналогічно як у попередньому критерії область прийняття нашої гіпотези набуває вигляду:

$$|t(n)| < t_{\frac{\alpha}{2}}(n-2),$$

де $t_{\frac{\alpha}{2}}(n-2)$ – $100\dfrac{\alpha}{2}$ відсоткова точка $t$-розподілу Стьюдента з $(n-2)$ степенями свободи, використання якої є більш зручнішим.

А далі, якщо вимір $x_{i(n)}$ виявився аномальним, то його вилучають з вибірки і всю процедуру повторюють з пункту 1, але вже зі скороченою вибіркою, до тих пір поки найбільш підозрілий на аномальність вимір виявиться не викидом.

**Самостійна робота №3.** З навчального посібника «Слабоспицький О.С. Аналіз даних. Попередня обробка, 2001» необхідно пропрацювати матеріал наведений у:
**Додаток 1.** Нормальний закон та пов'язані з ним розподіли.

**3) Критерій Тітьєна-Мура.** Є можливість перевірки на аномальність одразу $k\,(k \geq 1)$ найбільш підозрілих спостережень. Розв'язання такої задачі зводиться до перевірки гіпотези:

$H_0$ : *не усі k вимірів, найбільш підозрілі на аномальність, є викидами*, $\alpha > 0$.

---

1. Знову підраховуємо $\overline{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2. Далі будуємо послідовність:

$$z_1, z_2, \quad \ldots, \quad z_n,$$

де $z_i = \left| x_i - \overline{x}(n) \right|,\ i = \overline{1,n}$.

3. А потім відповідний варіаційний ряд:

$$z_{(1)}, z_{(2)}, \quad \ldots, \quad z_{(n-k)}, z_{(n-k+1)}, \quad \ldots, z_{(n)},$$

де $z_{(j)} = \left| x_{i(j)} - \overline{x}(n) \right|,\ j = \overline{1,n}$.

Найбільш підозрілими на аномальність будемо вважати $k$ вимірів $x_{i(j)}$, які відповідають $k$ останнім членам варіаційного ряду $z_{(1)}, z_{(2)}, \quad \ldots, \quad z_{(n)}$, а саме виміри: $x_{i(n-k+1)}, x_{i(n-k+2)}, \ldots, x_{i(n)}$.

4. Тепер обчислимо статистику:

$$E(n,k)=\frac{\sum_{i=1}^{n-k}\left(z_{(i)}-\overline{z}(n-k)\right)^2}{\sum_{i=1}^{n}\left(z_{(i)}-\overline{z}(n)\right)^2},\quad \text{де } \overline{z}(m)=\frac{1}{m}\sum_{i=1}^{m}z_{(i)}.$$

5. Логічно в якості *області прийняття нашої гіпотези* $H_0$, що спостереження $x_{i(n-k+1)}, x_{i(n-k+2)}, \ldots, x_{i(n)}$ не є викидами, взяти область, яка не включає у себе область малих значень статистики $E(n,k)$:

$$E(n,k) > E_\alpha(n,k),$$

де $E_\alpha(n,k)$ - квантиль рівня $\alpha$ розподілу статистики $E(n,k)$.

Якщо спостереження $x_{i(n-k+1)}, x_{i(n-k+2)}, \ldots, x_{i(n)}$ виявилися аномальними, то їх видаляють з вибірки і всю процедуру повторюють, але вже зі скороченою вибіркою, інакше алгоритм завершує свою роботу.

**Недоліки критерію Тітьєна-Мура.**

- не формалізована процедура вибору значення $k\,(k \ge 1)$ (можна запропонувати метод ділення навпіл: $k := k_0$; а потім $k := [k/2]$ тобто в результаті $k$ буде приймати таку послідовність значень $k_0, [k_0/2], [k_0/2^2], \ldots, 3, 2, 1$ (тут використовується критерій Граббса або Томпсона), де $k_0$ - стартове значення),
- критерій Тітьєна-Мура сильно залежить від нормальності.

**4) Засоби розвідувального аналізу** видалення аномальних скалярних спостережень розглядаються у відповідному розділі, який присвячений розвідувальному аналізу.

## II. *Випадок обробки векторних спостережень.*

**II.** ***Випадок обробки векторних спостережень.*** Припустимо тепер, що спостереження є векторними:

$$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n, \qquad \vec{x}_i \in \mathbb{R}^q, \ i = \overline{1,n}.$$

У векторному випадку можна скористатися такими засобами:

1) критерієм на базі $F$ - статистики,

2) діаграмою розсіювання (у випадку $q = 2$ ).

**1) Критерій на базі $F$ - статистики.** Задачу будемо розв`язувати шляхом перевірки гіпотези:

$H_0$ : *найбільш підозрілий на аномальність вимір не є викидом,* $\alpha > 0$.

1. Спочатку обчислюємо наступні величини:

$$\bar{\vec{x}}_i(n) = \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} \vec{x}_j, \quad \hat{V}_i(n) = \frac{1}{n-2}\sum_{\substack{j=1 \\ j\neq i}}^{n} \left(\vec{x}_j - \bar{\vec{x}}_i(n)\right)\left(\vec{x}_j - \bar{\vec{x}}_i(n)\right)^T, \quad i = \overline{1,n}.$$

Махаланобіса:

$$d_i^2(n) = \left(\vec{x}_i - \bar{\vec{x}}_i(n)\right)^T \hat{V}_i^{-1}(n)\left(\vec{x}_i - \bar{\vec{x}}_i(n)\right), \quad i = \overline{1,n}.$$

3. Знаходимо індекс найбільш підозрілого на аномальність виміру

$$i_0 = \arg\max_i d_i^2(n).$$

4. Визначаємо значення такої статистики:

$$F_{i_0}(n) = \frac{(n-1)(n-1-q)}{n(n-2)q} d_{i_0}^2(n).$$

З`ясувалося, що розподіл статистики $F_{i_0}(n)$ можна наблизити $F$ –розподілом з параметрами $q$ та $(n-1-q)$.

5. Тоді, так як *область відхилення гіпотези* $H_0$ - це область великих значень статистики $F_{i_0}(n)$, то гіпотезу $H_0$ будемо приймати, якщо справедлива наступна нерівність:

$$F_{i_0}(n) < F_\alpha(q, n-1-q),$$

де $F_\alpha(q, n-1-q)$ – $100\alpha$ відсоткова точка $F$ – розподілу з $q$ та $n-1-q$ степенями свободи.

Якщо вектор спостережень $\vec{x}_{i_0}$ виявився аномальним, то його вилучають з вибірки, і всю процедуру повторюють починаючи з пункту 1 до того часу, доки на деякому кроці найбільш підозрілий на аномальність вимір виявиться не викидом. Після цього процедуру завершують.

**2) Діаграма розсіювання.** Вона описана у розділі, який присвячений розвідувальному аналізу.

---

## Перевірка стохастичності вибірки
### (Tests for Randomness)

Нехай досліджується вибірка

$$x_1, x_2, \ldots, x_n, \quad n \in \mathbb{N}.$$

Перед обробкою вибірки має сенс впевнитися, що вона є випадковою (стохастичною), а не знаходиться під впливом деякого систематичного зміщення, наприклад: монотонного або періодичного.

Розв'язувати проблему будемо шляхом перевірки гіпотези

$$H_0: \quad \textit{вибірка є стохастичною}, \ \alpha > 0.$$

Для цього пропонується використовувати:

1) критерій серій на базі медіани вибірки,
2) критерій зростаючих та спадаючих серій,
3) критерій квадратів послідовних різниць (критерій Аббе),

Зупинимося детальніше на кожному з них.

для перевірки гіпотези $H_0$, причому альтернативна гіпотеза виглядає таким чином

$H_1$: *у вибірці наявне систематичне монотонне зміщення середнього.*

1. Спочатку по вибірці визначається оцінка медіани $\hat{x}_{med}$.

2. Потім під $i$ – им членом вибірки $x_1, x_2, \ldots, x_n$, який більше $\hat{x}_{med}$ ставиться плюс, а який менше $\hat{x}_{med}$ ставиться мінус. Виміри які дорівнюють $\hat{x}_{med}$ до уваги не приймаються. Тобто символи розставляються згідно

$$\begin{cases} +, \text{якщо } x_i > \hat{x}_{med}, \\ \sqcup, \text{якщо } x_i = \hat{x}_{med}, \\ -, \text{якщо } x_i < \hat{x}_{med}. \end{cases}$$

В результаті отримаємо деяку послідовності плюсів та мінусів

$$+--+-\ldots-+$$

**Означення**. *Серія* - це підпослідовність підряд розташованих однакових символів у послідовності.

3. Далі обчислюємо такі дві статистики:

$\nu(n)$ - загальну кількість серій у ній,

$\tau(n)$ - кількість членів у найдовшій серії.

4. Зрозуміло, що вибірка буде мати стохастичну природу, якщо довжина найдовшої серії $\tau(n)$ не занадто довга, а загальна кількість серій $\nu(n)$ не занадто мала. Причому відомо, що розподіл статистики $\nu(n)$ можна наблизити деяким *нормальним розподілом*, а $\tau(n)$ - деяким *пуасонівським розподілом*. Тоді *область прийняття гіпотези*:

$$\begin{cases} \nu(n) > \nu_\beta(n), \\ \tau(n) < \tau_{1-\beta}(n), \end{cases}$$

де $\nu_\beta(n), \tau_\beta(n)$ – квантилі рівня $\beta$ статистик $\nu(n)$ та $\tau(n)$ відповідно.

А $\beta$ вибирається таким чином, щоб помилка I роду не перевищувала $\alpha$. Відомо, що при фіксованому значенні $\beta$ рівень значущості $\alpha$ буде належати інтервалу $\left[\beta, 2\beta-\beta^2\right]$. Останнє дозволяє визначити $\beta$, як розв'язок такого співвідношення $2\beta-\beta^2 \le \alpha$.

**Самостійна робота.** Знайти значення $\beta$, як розв'язок останньої нерівності при заданому $\alpha$.

**2) Критерій зростаючих та спадаючих серій.** Перевірку гіпотези $H_0$ за допомогою цього критерію будемо здійснювати при умові, що його альтернативна гіпотеза виглядає таким чином

$H_1$: *у вибірці наявне систематичне монотонне або циклічне зміщення середнього.*

1. Спочатку у вибірці $x_1, x_2, \ldots, x_n$ замінюємо підряд розташовані однакові виміри одним їх представником.

$$x_1, x_2, \ldots, x_{n'}$$

ставимо символ плюс, якщо його наступний член з вибірки строго більше поточного, і ставимо символ мінус, якщо його наступний член з вибірки строго менше поточного, тобто символи під $i-$им членом розставляються згідно

$$\begin{cases} +, \text{ якщо } x'_i < x'_{i+1}, \\ -, \text{ якщо } x'_i > x'_{i+1}. \end{cases}$$

Отримуємо деяку послідовність довжини $(n'-1)$ плюсів та мінусів:

$$+-+-\ldots++$$

3. Далі, на базі утвореної послідовності плюсів та мінусів, визначаємо статистики $\nu(n)$ та $\tau(n)$ абсолютно аналогічно, як це робилося у попередньому критерії.

4. Потім використовується та ж сама ідея для побудови *області прийняття нашої гіпотези* про стохастичність вибірки. Вона буде мати ідентичний вигляд:

$$\begin{cases} \nu(n) > \nu_\beta(n), \\ \tau(n) < \tau_{1-\beta}(n), \end{cases}$$

де $\nu_\beta(n), \tau_\beta(n)$ – квантилі рівня $\beta$ статистик $\nu(n)$ та $\tau(n)$ відповідно.

Зауваження відносно вибору $\beta$ залишається у силі.

**3) Критерій квадратів послідовних різниць** (критерій Аббе). Даний критерій використовується при роботі *з нормальними вибірками*. *На цьому класі* вибірок він *є більш потужним* ніж попередні критерії.

Згадаємо як розуміти, що критерій $K_1$ є більш потужний ніж інший критерій $K_2$?

При перевірці гіпотези $H_0$ за допомогою цього критерія, у нього в якості альтернативної виступає така гіпотеза

$H_1$: *у вибірці наявне систематичне зміщення середнього.*

1. На основі вибірки підрахуємо таку статистику:

$$\gamma(n) = \frac{\frac{1}{2(n-1)}\sum_{i=1}^{n-1}(x_{i+1}-x_i)^2}{\frac{1}{n-1}\left[\left(\sum_{i=1}^{n}x_i^2\right) - n\bar{x}^2(n)\right]},$$

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

2. Легко бачити, що до області відхилення гіпотези $H_0$ потрібно віднести область малих значень, тоді *область прийняття гіпотези* буде мати вигляд:

$$\gamma(n) > \gamma_\alpha(n),$$

де $\gamma_\alpha(n)$ – квантиль рівня $\alpha$ статистики $\gamma(n)$, який можна підрахувати таким чином

$$\gamma_\alpha(n) = \begin{cases} \text{визначається по табл. 4.9 з роботи } [*], \text{ якщо } n \le 60, \\ 1 + \dfrac{u_\alpha}{\left[ n + \dfrac{1}{2}\left(1 + u_\alpha^2\right) \right]^{\frac{1}{2}}}, \text{ якщо } n > 60, \end{cases}$$

а $u_\alpha$ – квантиль рівня $\alpha$ стандартного нормального розподілу,

[*] Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983.

*Практика використання* критеріїв перевірки вибірки на стохастичность. …

## Розвідувальний аналіз (Exploratory data analysis)

**Розвідувальний аналіз** – це один із етапів попередньої обробки даних, який дозволяє провести візуальний експрес-аналіз даних на основі засобів їх візуалізації або перетворення, тобто представлення даних у зручному для оперативного аналізу вигляді.

Наприклад, у вигляді різноманітних графіків, діаграм, схем, таблиць і т.п. Результати цього аналізу слугуватимуть відправною точкою для планування подальшої поглибленої обробки інформації.

**I.** ***Випадок обробки скалярних спостережень.*** У цій ситуації у розвідувальному аналізі можна використовувати:

1) пробіт-графік (probit plot),
2) ймовірнісний графік (probability plot),
3) висячі гістобари (hanging histobars),
4) завмерлу коренеграму (suspended rootogram),
5) зображення "скринька з вусами" (box-and-whisker plot) та його модифікації
   (multiple box-and-whisker plot, notched box-and-whisker plot),
6) зображення "стебло-листок" (stem-and-leaf plot), і т.д.

**II.** ***Випадок обробки двовимірних спостережень.*** У цій ситуації у розвідувальному аналізі можна використовувати:

1) діаграму розсіювання (scatter diagram),
2) таблиця спряженості (contingency table).

Далі буде розглянуто послідовно можливості кожного з цих засобів.

## Класи розподілів типу зсув-масштабу

Для опису перших двох графічних представлень даних потрібно познайомитися з класами розподілів типу зсув-масштабу.

**Означення.** Клас розподілів $\mathcal{F}$ називається ***класом розподілів типу зсув-масштабу***, якщо існує така базова функція розподілу $F_0(\cdot) \in \mathcal{F}$, що для будь-якої функції розподілу $F(\cdot)$ з цього класу існують дійсні $a$ та $b$ $(b > 0)$ такі, що її можна представити таким чином:

$$F(x) = F_0\left(\frac{x-a}{b}\right).$$



Зауважимо, що параметр $a$ називають ***параметром зсуву***, а $b$ - ***параметром масштабу***.

Приклади класів розподілів типу зсув-масштабу.

1. Клас нормальних розподілів.

Дійсно довільну функцію розподілу $F(x)$ нормально розподіленої величини $\xi \sim \mathcal{N}\left(m,\sigma^2\right)$ можна представити у вигляді

$$F(x)=\Phi\left(\frac{x-m}{\sigma}\right),$$

де $\Phi(x)$ – функція розподілу нормально розподіленої величини з параметрами 0 та 1. Для цього класу розподілів: $\Phi(\cdot)$ – базова функція, $a=m,\ b=\sigma$.

2. Клас показникових (експоненціальних) розподілів.

Нехай $F(x)$ - функція показникового розподілу з параметром $\lambda$ $(\lambda>0)$, тобто

$$p(x)=F'(x)=\begin{cases}\lambda e^{-\lambda x}, \text{ якщо } x\geq 0,\\ 0 \quad , \text{ якщо } x<0.\end{cases}$$

Тоді справедливо

$$F(x)=\Phi_1(\lambda x),$$

де $\Phi_1(x)$ – функція експоненціального розподілу з параметром 1. Тобто роль базової функції тут відіграє функція $\Phi_1(\cdot)$, а потрібні константи визначаються згідно з $a=0,\ b=\lambda^{-1}$.

## I. Випадок обробки скалярних спостережень

Детальніше:

**1) Пробіт-графік (probit plot).**

Нехай $\mathcal{F}$ - деякий клас розподілів типу зсув-масштабу з базовою функцією $F_0(\cdot)$ для якої існує $F_0^{-1}(\cdot)$.

Розглянемо обробку вибірки спостережень $x_1, x_2, \ldots, x_n$ над скалярною змінною $\xi$ з функцією розподілу $F_\xi(x)$.

Подивимося, який повинен мати вигляд побудований пробіт-графік у випадку коли функція розподілу випадкової величини $\xi$, яка спостерігається, належить цьому класу розподілів $\mathcal{F}$. Тоді існують $a$ та $b$ $(b>0)$ такі, що

$$\hat{F}_\xi(x) \approx F_\xi(x) = F_0\left(\frac{x-a}{b}\right).$$

А сам пробіт-графік буде мати такий вигляд:

$$y = F_0^{-1}(\hat{F}_\xi(x)) \approx F_0^{-1}\left(F_0\left(\frac{x-a}{b}\right)\right) = \frac{x}{b} - \frac{a}{b}.$$

*Призначення*. Це дозволяє використовувати цей графік для візуального розв'язку наступних задач:

1) *Перевірки гіпотези* $H_0 : F_\xi(\cdot) \in \mathcal{F}$.

   У випадку справедливості цієї гіпотези пробіт-графік буде уявляти собою приблизно деяку пряму, в протилежному випадку гіпотезу відхиляють.

2) *Виявлення наявності аномальних спостережень у вибірці.* Про присутність викидів у вибірці буде говорити наявність деяких точок графіку, які розташовані суттєво осторонь основної маси точок графіку.

**2) Ймовірнісний графік (probability plot).**

*Побудова.* Нехай $\hat{F}_\xi(x)$ – емпірична функція розподілу, яка обчислена по вибірці спостережень $x_1, x_2, \ldots, x_n$ над випадковою величиною $\xi$.

***Ймовірнісний графік для класу розподілів*** $\mathcal{F}$ – це графік функції $y = \hat{F}_\xi(x)$, побудований на спеціальному *ймовірнісному папері класу розподілів* $\mathcal{F}$. Останній відрізняється від звичайного паперу зміненим масштабом по осі $y$. З цією метою на такому папері смугу $\{(x,y) : 0 \le y \le 1\}$ трансформують таким чином: $(x,y) \to (x, F_0^{-1}(y))$.

*Призначення*. Можливості та методика використання ймовірнісного графіку точно такі як і у пробіт-графіку:

1) *Перевірки гіпотези* $H_0 : F_\xi(\cdot) \in \mathcal{F}$.

У випадку справедливості цієї гіпотези пробіт-графік буде уявляти собою приблизно деяку пряму, в протилежному випадку гіпотезу відхиляють.

2) *Виявлення наявності аномальних спостережень у вибірці.* Про присутність викидів у вибірці буде говорити наявність деяких точок графіку, які розташовані суттєво осторонь основної маси точок графіку.

Якщо $\mathcal{F}$ – клас нормальних розподілів, то цей графік називають *нормальним ймовірнісним графіком*, а відповідний папір – *нормальним ймовірнісним папером*.

Наступні два засоби розвідувального аналізу торкаються візульної перевірки гіпотези нормальності.

Наступні два засоби розвідувального аналізу торкаються візульної перевірки гіпотези нормальності.

**Означення.** ***Нормальним розподілом найбільш узгодженим з вибіркою*** $x_1, x_2, \ldots, x_n$ називається такий нормальний закон

$$\mathcal{N}(\bar{x}(n), s^2(n)),$$

де $\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}(n)\right)^2$.

**3) Висячі гістобари (hanging histobars).**

*Побудова*. Спочатку по вибірці $x_1, x_2, \ldots, x_n$ визначають вибіркові значення математичного сподівання $\bar{x}(n)$ та дисперсії $s^2(n)$. Потім будується графік щільності нормального розподілу найбільш узгодженого з вибіркою $x_1, x_2, \ldots, x_n$, а саме $\mathcal{N}(\bar{x}(n), s^2(n))$. Далі у центрі кожного інтервалу групування даних до цієї кривої підвішують гістобару (вузенький прямокутник), висота якої пропорційна відносній частоті попадання вимірів у цей інтервал групування.

*Призначення*. Висячі гістобари використовують для візуальної перевірки гіпотези нормальності розподілу випадкової величини, яка спостерігається. Гіпотезу приймають, якщо основи гістобар незначно відхиляються від осі абсцис. В протилежному випадку її відхиляють.

**4) Завмерла коренеграма (suspended rootogram).**

*Побудова*. Вона представляє собою послідовність прямокутників, побудованих у центрах інтервалів групування даних вибірки $x_1, x_2, \ldots, x_n$, причому висота такого прямокутника для $i$-того інтервала групування даних пропорційна різниці

$$\sqrt{\nu_i^{(e)}} - \sqrt{\nu_i^{(t)}},$$

де $\nu_i^{(e)}, \nu_i^{(t)}$ – відповідно емпірична та теоретична відносні частоти попадання у $i$-тий інтервал групування. Остання підрахована згідно нормального розподілу найбільш узгодженого з вибіркою $x_1, x_2, \ldots, x_n$, тобто $\mathcal{N}(\bar{x}(n), s^2(n))$. Якщо $\hat{p}(x)$ - щільність нормального розподілу найбільш узгодженого з вибіркою $x_1, x_2, \ldots, x_n$, то

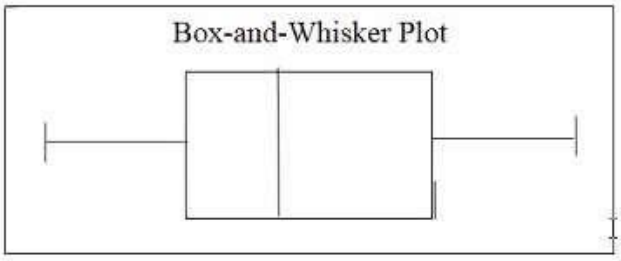
$$\nu_i^{(t)} = \int_{b_{i-1}}^{b_i} \hat{p}(x)dx,$$

де $b_{i-1}, b_i$ - лівий та правий кінці $i$-того інтервалу групування.

*Призначення*. Це графічне представлення можна використовувати для візуальної перевірки гіпотези про нормальність розподілу випадкової величини, яка спостерігається. Остання попередньо вважається нормально розподіленою, якщо побудовані прямокутники незначно відхиляються від осі абсцис, інакше вона відхиляється.

**5) Зображення "скринька з вусами" (box-and-whisker plot).**

*Побудова*. Воно має у загальному випадку такий нижченаведений вигляд:




*Модифікації*: multiple box-and-whisker plot, notched box-and-whisker plot.

*Призначення*. Це зображення надає можливість отримати таку інформацію. Проекція середньої вертикальної лінії скриньки на вісь абсцис дає нам значення медіани, лівої границі скриньки – нижнього квартилю, правої границі скриньки – верхнього квартилю. Проекції лівого кінця лівого вуса та правого кінця правого вуса відповідно дають нам найменше найбільше значення у вибірці. При наявності у вибірці викидів (вимірів, які знаходяться від скриньки на відстані більшій ніж півтори інтерквартильної широти), на зображенні вони будуть представлені у вигляді окремих точок, відображених лівіше та правіше кінців вищевказаних ліній.

Stem-and-Leaf Plot for Variable1: unit = 100  1|2 represents 1200

| | | |
|---|---|---|
| | LO | 18, 19, 21, 21 |
| 7 | 2F | 455 |
| 18 | 2S | 66666777777 |
| 30 | 2° | 888888999999 |
| 46 | 3* | 00000000001111111 |
| 66 | 3T | 222222233333333333333 |
| 84 | 3F | 4444444445555555555 |
| 106 | 3S | 66666666667777777777777 |
| 73 | 3° | 88888888888999999999999999 |
| 47 | 4* | 000000000011111 |
| 32 | 4T | 22333 |
| 27 | 4F | 444444455555555 |
| 12 | 4S | 6666777 |
| 5 | 4° | 89 |
| 3 | 5* | 0 |
| | HI | 61, 73 |

У першому рядку вказано, що це зображення побудовано для змінної Variable1, використовуючи масштабний множник 100. Лівіше вертикальної риски вказується ведуча цифра поточного виміру з одним із фіксованих символів, а правіше вертикальної риски його наступна цифра. Врахувавши масштабний множник 100, одразу отримаємо значення поточного спостереження. Цифра, яка стоїть у першому стовпчику вказує кількість відображених спостережень у поточному рядку плюс у всіх рядках до найближчого краю зображення. У самих крайніх рядках, які починаються з абревіатур LO або HI, можуть вказуватися виміри підозрілі на аномальність.

*Призначення*. Зображення "стебло-листок" дозволяє візуально з'ясувати загальний вигляд розподілу даних, інтервал їх концентрації, симетричність розподілу, наявність вимірів підозрих на аномальність.

## II. Випадок обробки двовимірних спостережень

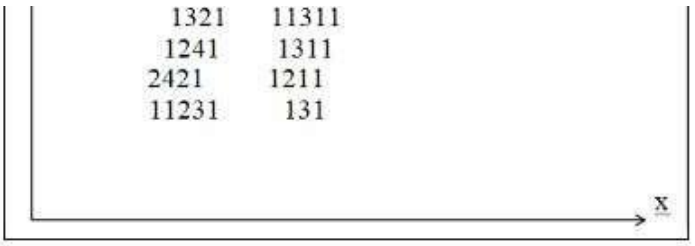
**1) Діаграма розсіювання (scatter diagram).**

*Побудова*. Нехай маємо спостереження над двома кількісними скалярними змінними

$$\xi : x_1, x_2, \ldots, x_n,$$

$$\eta : y_1, y_2, \ldots, y_n.$$

Спочатку представимо ці виміри на екрані монітора, який працює у текстовому режимі. У цій ситуації весь екран монітора розбивається на знакомісця (прямокутники). Підраховуємо скільки значень пар $(x_i, y_i)$ з вибірки, тобто точок з координатами $(x_i, y_i)$ попало у кожне знакомісце. А потім усі ці ненульові значення виводимо у відповідні знакомісця. Якщо ці значення з другого десятка, то можна використовувати режим «інверсії», а якщо з третього – режим «blink». У підсумку, побачимо щось на зразок такого




Якщо монітор працює у графічному режимі, то зображення на екрані формується за допомогою різнокольорових пікселів. Тут вже підраховуємо скільки точок з координатами $(x_i, y_i)$ попало в площину кожного пікселя. А потім кожен піксель виводимо на екран тим темнішим відтінком коричневого, чим більше значень пар $(x_i, y_i)$ з вибірки попало в площину цього пікселя.

*Призначення*. Діаграма розсіювання дозволяє таке:

- з`ясувати загальний вигляд залежності (класу функцій апроксимації залежності) між $\xi$ та $\eta$,
- з`ясувати наявність аномальних спостережень.

**2) Таблиця спряженості (contingency table).**

*Призначення*. Використовується для табличного представлення спостережень над двома скалярними змінними зі скінченними множинами значень. Це можуть бути номінальні, ординальні, кількісні дискретні або кількісні неперервні змінні, спостереження над якими згруповані.

*Побудова*. Нехай змінна $\eta$ приймає всього $m_1$ значень, а змінна $\xi$- $m_2$ значень. Вважаємо, що отримали $n$ спостережень над цими двома скалярними змінними. Тоді кількість наслідків спостережень $n_{ij}$, коли змінна $\eta$ прийняла своє $i-$те значення, а $\xi$ своє $j-$те

значення заносимо у комірку на перетені $i$ – того рядка та $j$ – того стовпчика такої таблиці, яку і називають ***таблицею спряженості***:

| η \ ξ | 1 | 2 | • • • | $j$ | ⋯ | $m_2$ | $\sum$ |
|---|---|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | • • • | $n_{1j}$ | ⋯ | $n_{1m_2}$ | $n_{1\bullet}$ |
| 2 | $n_{21}$ | $n_{22}$ | • • • | $n_{2j}$ | ⋯ | $n_{2m_2}$ | $n_{2\bullet}$ |
| ⋮ | ⋮ | ⋮ | • • • | ⋮ | ⋯ | ⋮ | ⋮ |
| $i$ | $n_{i1}$ | $n_{i2}$ | • • • | $n_{ij}$ | ⋯ | $n_{im_2}$ | $n_{i\bullet}$ |
| ⋮ | ⋮ | ⋮ | • • • | ⋮ | ⋯ | ⋮ | ⋮ |
| $m_1$ | $n_{m_1 1}$ | $n_{m_1 2}$ | • • • | $n_{m_1 j}$ | ⋯ | $n_{m_1 m_2}$ | $n_{m_1\bullet}$ |
| $\sum$ | $n_{\bullet 1}$ | $n_{\bullet 2}$ | • • • | $n_{\bullet j}$ | ⋯ | $n_{\bullet m_2}$ | $n$ |

Значення $n_{ij}$ називають *частотою відповідної комірки*. Значення у останньому рядку – це сума по стовпчикам, значення у останньому стовпчику – це сума по рядкам таблиці спряженості, а значення у правому нижньому кутку – це загальна кількість спостережень, тобто:

$$n_{i\bullet} = \sum_{j=1}^{m_2} n_{ij}\ ,\ n_{\bullet j} = \sum_{i=1}^{m_1} n_{ij}\ ,\ n = \sum_{i=1}^{m_1} n_{i\bullet} = \sum_{j=1}^{m_2} n_{\bullet j}\ .$$

Тут крапка у позначеннях замість індексу позначає, що було здійснено сумування по тому індексу замість якого вона фігурує.



### 1.3. Моделювання змінних

Під час роботи з даними виникає потреба в генерації спостережень над випадковими величинами з заданими функціями розподілу. Отримати такі вибірки можна шляхом моделювання потрібних значень. Один із підходів при розв'язанні цієї задачі полягає у пред-

ставленні величини $\xi$, яка моделюється, у вигляді деякої функції $g(\xi_1,\xi_2,\ldots,\xi_q)$ від найпростіших випадкових величин $\xi_1,\xi_2,\ldots,\xi_q$. Як правило, у цій ролі виступають незалежні $\xi_1,\xi_2,\ldots,\xi_q$, рівномірно розподілені на відрізку $[0,1)$. Тому, задача зводиться до необхідності вміти розв'язувати наступні дві проблеми:

- моделювання незалежних $\xi_1,\xi_2,\ldots,\xi_q$, рівномірно розподілених на відрізку $[0,1)$,
- знаходження потрібної функції $g(\cdot,\cdot,\ldots,\cdot)$ у представленні величини $\xi$, яку потрібно моделювати.

Нижче кожну з них розглянемо окремо.

Перша проблема розв'язується шляхом використання *датчика (генератора) випадкових чисел* – спеціального пристрою, який після запиту на виході дозволяє отримати реалізацію випадкової величини з заданим законом розподілу. Повторні звернення до цього генератора дозволяють отримати наступні незалежні спостереження над цією випадковою величиною. Найбільший інтерес для нас представляє датчик рівномірно розподіленої на відрізку $[0,1)$ випадкової величини. Познайомимося з генераторами випадкових чисел детальніше.

### 1.3.1. Класифікація датчиків випадкових чисел

Виділяють наступні *класи датчиків* (генераторів) *випадкових чисел*:

- табличні,
- фізичні,
- програмні.

Охарактеризуємо кожен із цих типів генераторів окремо.

*Табличний датчик випадкових чисел* являє собою таблицю заповнену реалізаціями випадкової величини з заданим законом розподілу. Представлені у таких таблицях вибірки, як правило, досить високої якості, але вони мають обмежений об'єм. Та й кількість таких вибірок невелика. Це суттєво стримує їх використання.

*Фізичний датчик випадкових чисел* конструюється на основі деякого електронного пристрою, на виході якого спостерігають потрібну реалізацію. Ці генератори надають можливість отримувати ви-

бірки довільного об'єму, що не дозволяли робити табличні датчики. Але вони мають інший недолік – кожна отримана вибірка є унікальною і повторити її практично неможливо. Цього недоліку позбавлений наступний клас генераторів.

*Програмний датчик випадкових чисел* будується на базі деякої програми, на виході якої формується потрібна реалізація. Ці програми, як правило, базуються на використанні певних рекурентних формул із деякою глибиною пам'яті. Задаючи у рекурентному співвідношенні однакові початкові значення (стартові числа), можна повторити конкретну вибірку довільну кількість раз. Але ці генератори, в свою чергу, мають свій недолік – вони періодичні. В принципі, числа, які отримуються на виході, потрібно називати "псевдовипадковими" числами, бо вони формуються згідно з детермінованого закону, а саме деякою рекурентною формулою.

### 1.3.2. Програмні датчики та їх властивості

В основі програмних генераторів, як правило, лежить використання рекурентних формул. Саме вибір їх конкретного вигляду і буде визначати властивості цих датчиків. Познайомимося з деякими лінійними, а потім нелінійними формулами, які знайшли застосування в генераторах рівномірно розподілених на відрізку $[0,1)$ випадкових величин.

Широкого розповсюдження при побудові цих датчиків набула *лінійна змішана формула*. Вона має наступний загальний вигляд:

$$\begin{cases} x_i = \dfrac{\widetilde{x}_i}{M}, \\ \widetilde{x}_i = \left( a_0 + \sum\limits_{j=1}^{l} a_j \widetilde{x}_{i-j} \right) \bmod M, \quad i = 1, 2, \ldots \end{cases},$$

де всі параметри алгоритму є цілими і задовольняють таким умовам: $l \geq 1$, $a_j \geq 0$ $(j = \overline{0,l})$, $M > 0$, а стартові числа лежать у межах $0 \leq \widetilde{x}_{1-j} \leq M-1$, $(j = \overline{1,l})$.

Так як по побудові послідовність $\{\tilde{x}_i\}_{i>0}$ набуває значення з множини $\{0, 1, \ldots, M-1\}$, то послідовність $\{x_i\}_{i>0}$ в свою чергу буде набувати значення з потрібного інтервалу $[0,1)$. Зауважимо, що значення $\{\tilde{x}_i\}_{i\geq 0}$ можна використовувати при моделюванні рівномірного розподілу на множині значень $\{0, 1, \ldots, M-1\}$.

Аналіз лінійної змішаної формули почнемо з деяких її частинних випадків. Розглянемо випадок, коли $a_0 = 0$ та $l = 1$. Кажуть, що у цьому випадку використовується *мультиплікативний конгруентний метод*, а сам алгоритм набуває вигляд:

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M}, \\ \tilde{x}_i = (a_1 \tilde{x}_{i-1}) \bmod M, \quad i = 1, 2, \ldots \end{cases}.$$

Для того, щоб скористатися цією процедурою достатньо знати одне стартове число $\tilde{x}_0$ $(0 \leq \tilde{x}_0 \leq M-1)$. Очевидно, що вибір $\tilde{x}_0 = 0$ буде невдалим, бо тоді вся послідовність $\{\tilde{x}_i\}_{i\geq 0}$ буде тотожна нулеві і така вибірка не представляє цінності.

З алгоритму випливає, що можливими значеннями у послідовності $\{\tilde{x}_i\}_{i>0}$ є значення з множини $\{0, 1, \ldots, M-1\}$, а сама вона буде періодичною. Причому максимальне значення періоду $T_{max}$ не буде перевищувати $M$, а це в свою чергу примушує вибирати $M$ якомога ближчим до максимального цілого, що допускається на конкретному комп'ютері, і яке будемо позначати у подальшому через $\max\text{int}$. Наприклад, як $M$ можемо взяти найбільше просте число, яке менше $\max\text{int}$. Таким чином, виникає потреба у виборі параметрів алгоритму мультиплікативного конгруентного методу так, щоб максимізувати його період. А так як період довжини $M$ повинен містити у собі значення рівне нулеві, то це приводить до подальшого виродження послідовності $\{\tilde{x}_i\}_{i\geq 0}$ у нуль, що приводить до висновку, що мультиплікативний конгруентний метод не дозволяє досягти максимального теоретично можливого періоду рівного $M$.

Визначимо функцію $\lambda(M)$ наступним чином:

$$\lambda(M)=\begin{cases}1, & \text{якщо } M=2,\\ 2, & \text{якщо } M=4,\\ p^{q-1}(p-1), & \text{якщо } M=p^q \quad (q\ge 1, \text{просте } p>2),\\ НСК(\lambda(p_1^{q_1}),\lambda(p_2^{q_2}),\ldots,\lambda(p_k^{q_k})), & \text{якщо } M=p_1^{q_1}p_2^{q_2}\cdots p_k^{q_k}\\ & (q_i\ge 1, \text{прості } p_i>2, i=\overline{1,k}),\end{cases}$$

де $НСК(n_1,n_2,\ldots,n_k)$ – найменше спільне кратне для позитивних цілих чисел $n_1,n_2,\ldots,n_k$.

Наступне відоме твердження дозволяє з'ясувати питання про максимально можливий період досліджуємого методу.

**Теорема 1**. Максимальний період послідовності $\{\tilde{x}_i\}_{i\ge 0}$ мультиплікативного конгруентного методу $T_{max}=\lambda(M)$. Для того, щоб він досягався достатньо виконання наступних умов:

1) $\tilde{x}_0$ та $M$ є взаємно прості числа,

2) $a_1^{\lambda(M)} \bmod M=1$, тобто $a_1$ є первісним елементом по модулю $M$.

*Зауваження*. Якщо покласти $M$ рівним деякому простому числу, то $T_{max}=M-1$. Тобто у цьому випадку період буде тільки на одиницю менший від $M$. Наприклад, можна взяти найбільше ціле, яке менше max int. Тоді залежно від розрядності комп'ютера $q$ можна скористатися наступними значеннями для $M$:

| $q$ | $M$ |
|---|---|
| 16 | $2^{16}$-15 |
| 32 | $2^{32}$-5 |
| 64 | $2^{64}$-59 |

А використання у ролі $M$ значення $2^q$ $(q\ge 3)$ дозволяє досягти лише періоду $M/4$.

Перейдемо тепер до розгляду випадку, коли $a_0>0$ та $p=1$. Метод, який йому відповідає, називається *змішаним конгруентним методом*, а безпосередньо алгоритм має вигляд:

$$\begin{cases}x_i=\dfrac{\tilde{x}_i}{M},\\ \tilde{x}_i=(a_0+a_1\tilde{x}_{i-1}) \bmod M, \quad i=1,2,\ldots .\end{cases}$$

Вибір параметрів цієї процедури з метою досягнення максимального періоду можна здійснити скориставшись нижченаведеним твердженням.

**Теорема 2**. Для того, щоб визначена згідно змішаного конгруентного методу послідовність $\{\tilde{x}_i\}_{i\geq 0}$ досягала свого максимального періоду $T_{\max} = M$ необхідно і достатньо, щоб виконувалися наступні умови:

1) $a_0$ і $M$ – взаємно прості,

2) $(a_1 - 1) \bmod p = 0$ для кожного простого $p$, яке є дільником $M$,

3) $(a_1 - 1) \bmod 4 = 0$, якщо $M \bmod 4 = 0$.

*Зауваження*. Вибір параметрів алгоритму змішаного конгруентного методу, які забезпечують його максимальний період, не може бути гарантією високої якості побудованого датчика випадкових чисел. Про це красномовно свідчить наступний приклад генератора:

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M}, \\ \tilde{x}_i = (1 + \tilde{x}_{i-1}) \bmod M, \quad i = 1,2,\ldots \end{cases}.$$

Легко бачити, що він має максимальний період, але послідовність $\{\tilde{x}_i\}_{i\geq 0}$ далека від випадкової. Тому для цього потрібно проводити додаткові дослідження.

Перехід від лінійної функції до квадратичної приводить до *квадратичного конгруентного методу:*

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M}, \\ \tilde{x}_i = (a_0 + a_1\tilde{x}_{i-1} + a_2\tilde{x}_{i-1}^2) \bmod M, \quad i = 1,2,\ldots \end{cases};$$

який, очевидно, залишає значення максимального періоду без змін: $T_{\max} = M$.

Подальше збільшення максимального періоду можна досягти, зробивши у лінійній змішаній формулі датчика, яка використовується, глибину пам'яті $l > 1$:

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M}, \\ \tilde{x}_i = \left(a_0 + \sum_{j=1}^{l} a_j \tilde{x}_{i-j}\right) \bmod M, \quad i = 1,2,\ldots \end{cases}.$$

Або взагалі звернутися до конструкції генератора наступного загального вигляду з деякою функцією $g(\cdot,\cdot,\ldots,\cdot)$ від $l\,(l>1)$ попередніх значень $\tilde{x}_{i-1}, \tilde{x}_{i-2}, \ldots, \tilde{x}_{i-l}$:

$$\begin{cases} x_i = \dfrac{\tilde{x}_i}{M}, \\ \tilde{x}_i = g(\tilde{x}_{i-1}, \tilde{x}_{i-2}, \ldots, \tilde{x}_{i-l}), \quad 0 \le \tilde{x}_i \le M-1, \quad i = 1,2,\ldots \end{cases}.$$

Останнє ускладнення структури датчика приводить до того, що можливий найбільший період може досягати значення $T_{\max} = M^l$.

### 1.3.3. Моделювання дискретних випадкових величин

Скористаємося побудованими датчиками рівномірно розподіленої на відрізку $[0,1)$ випадкової величини для моделювання деяких стохастичних змінних. Почнемо з моделювання дискретної випадкової величини $\xi$. Нехай вона приймає значення $y_i$ з ймовірностями $p_i = P\{\xi = i\}$, $i = \overline{1,m}$.

Так як $\sum_{i=1}^{m} p_i = 1$, то інтервал $[0,1)$ можна розбити на $m$ підінтервалів:

$$\Delta_1 = [0, p_1), \Delta_2 = [p_1, p_1 + p_2), \ldots, \Delta_i = \left[\sum_{j=1}^{i-1} p_j, \sum_{j=1}^{i} p_j\right), \ldots, \Delta_m = \left[\sum_{j=1}^{m-1} p_j, 1\right)$$

причому довжина інтервалу $\Delta_i$ буде дорівнювати $p_i$ $(i = \overline{1,m})$. Це дозволяє запропонувати наступний простий алгоритм моделювання дискретної випадкової величини $\xi$.

Звертаємося до датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової величини і отримуємо на його виході деяке значення $x$, яке попаде в один із побудованих підінтервалів. Тоді, якщо $x \in \Delta_i$, то логічно вважати що $\xi$ прийняла значення $y_i$. Повторивши цю процедуру необхідну кількість раз, отримаємо вибірку потрібного об'єму. (У всіх наведених далі алгоритмах моделювання випадкових величин буде описано лише перший крок.)

Цей алгоритм без проблем переноситься на випадок, коли потрібно моделювати дискретну випадкову величину $\xi$, яка має злічену множину значень $y_i$, що приймаються з відповідними ймовірностями $p_i$, $i = 1, 2, 3, \ldots$. Наявність рекурентних співвідношень для величин $p_i$ може суттєво спростити процедуру моделювання. Особливо це корисно у нашому випадку зліченої множини значень $\xi$. Для ряду відомих розподілів нескладно виписати потрібні рекуренти.

Приклади. 1. Якщо випадкова величина $\xi$ має *геометричний розподіл* з параметром $p$ $(0 < p < 1)$, то для ймовірностей $p_i = p(1-p)^i$ $(i \geq 0)$ очевидно справедливо $p_{i+1} = p_i(1-p)$, $p_0 = p$, $i \geq 0$.

2. Для *розподілу Пуасона* з параметром $\lambda$ $(\lambda > 0)$ у якого $p_i = \frac{\lambda^i}{i!}e^{-\lambda}$ $(i \geq 0)$ відповідний рекурент має вигляд $p_{i+1} = p_i \frac{\lambda}{i+1}$, $p_0 = e^{-\lambda}$, $i \geq 0$.

Простіший шлях можна запропонувати для моделювання рівномірного дискретного розподілу з $p_i = P\{\xi = i\} = \frac{1}{m}$, $i = \overline{1,m}$. Якщо $x$ – вихід датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової величини, то значення потрібної величини отримується по формулі $[1 + mx]$, де $[a]$ – ціла частина від $a$.

### 1.3.4. Моделювання неперервних випадкових величин

Перейдемо тепер до аналізу неперервного випадку. Нехай потрібно моделювати неперервну випадкову величину $\xi$. Позначимо її функцію розподілу через $F(z)$.

Розглянемо випадок коли $F(z)$ – строго монотонна функція. Тоді у ролі реалізації $\xi$ може виступити $F^{-1}(x)$, де $x$ – значення отримане з датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової

величини, а $F^{-1}(\cdot)$ - функція обернена до $F(x)$. Впевнимося у цьому.

Нехай $\eta$ – рівномірно розподілена на відрізку $[0,1)$ випадкова величина. Проаналізуємо функцію розподілу величини $F^{-1}(\eta)$:

$$P\{F^{-1}(\eta) < x\} = P\{\eta < F(x)\} = F(x).$$

Що і треба було довести.

Приклад. Застосуємо останній підхід до моделювання випадкової величини $\xi$, яка має *показниковий (експоненціальний) розподіл* з параметром $\lambda > 0$:

$$F(z) = \begin{cases} 1 - e^{-\lambda z}, & \text{якщо } z \geq 0, \\ 0, & \text{якщо } z < 0. \end{cases}$$

Дійсно, так як обернена функція має вигляд $F^{-1}(y) = -\dfrac{\ln(1-y)}{\lambda}$, то $-\dfrac{\ln(1-\eta)}{\lambda}$ має потрібний показниковий розподіл, де $\eta$ – величина рівномірно розподілена на інтервалі $[0,1)$. А так як $1-\eta$ теж рівномірно розподілена на інтервалі $[0.1)$. то можна зробити висновок, що величина $-\dfrac{\ln(\eta)}{\lambda}$ має показниковий розподіл із параметром $\lambda > 0$. Тоді у ролі реалізації $\xi$ може виступити $-\dfrac{\ln(x)}{\lambda}$, де $x$ – значення отримане з датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової величини.

Звернемося до моделювання *нормального розподілу* з параметрами $m$ та $\sigma^2$. Для цього скористаємося наступним твердженням.

**Теорема**. Нехай величини $\eta_1, \eta_2$ - незалежні, рівномірно розподілені на інтервалі $[0,1)$. Тоді випадкові величини

$$\xi_1 = \sin(2\pi\eta_1)\sqrt{-2\ln(\eta_2)},$$
$$\xi_2 = \cos(2\pi\eta_1)\sqrt{-2\ln(\eta_2)}$$

незалежні, нормально розподілені з параметрами 0 та 1.

Позначимо $x_1, x_2$ – незалежні спостереження над рівномірно розподіленою на інтервалі $[0,1)$ величиною. Тоді згідно теореми можна стверджувати, що значення

$$m + \sigma\sin(2\pi x_1)\sqrt{-2\ln(x_2)}, \quad m + \sigma\cos(2\pi x_1)\sqrt{-2\ln(x_2)}$$

є спостереженнями над незалежними, нормально розподіленими з параметрами $m$ та $\sigma^2$ величинами.

У разі необхідності моделювання випадкової величини *рівномірно розподіленої на інтервалі* $[a,b)$, достатньо скористатися очевидним перетворенням виходу $x$ датчика рівномірно розподіленої на відрізку $[0,1)$ випадкової величини: $a + (b-a)x$.