7. ЗАСТОСУВАННЯ ЧАСОВИХ РЯДІВ

7.1. Прогнозування

Для побудови прогнозу необхідно мати достатній для прояву статистичних закономірностей обсяг інформації. Зокрема рекомендується, щоб для річних спостережень було не менше, ніж п'ять точок, а для врахування сезонності – не менше трьох циклів. Іншою обов'язковою вимогою є забезпечення зіставлюваності даних. На основі змістового аналізу проблеми необхідно обґрунтувати можливість перенесення попередніх закономірностей на майбутній період. При невиконанні цих умов застосування математичних методів прогнозування не має сенсу.

Загальна схема прогнозування є такою. На першому етапі формулюють мету та завдання дослідження, а також підбирають необхідні статистичні дані. Здійснюють змістовий аналіз досліджуваного процесу й обирають показник (показники), що найбільш повно його характеризують, виявляють фактори, які істотно впливають на процес і визначають оптимальний **горизонт прогнозування** (термін, на який робиться прогноз). Останній обирається з урахуванням стабільності досліджуваного показника і, як правило, має бути не більшим, ніж 1/3 довжини наявного часового ряду.

На другому етапі здійснюють попередній аналіз наявних даних, зокрема перевіряють їх повноту, зіставлюваність, однорідність та стійкість. Після цього розраховують основні динамічні характеристики (прирости, темпи зростання, коефіцієнти автокореляції тощо) і будують графіки динаміки.

Наступним етапом є формування набору моделей досліджуваних динамічних рядів, їх ідентифікація й оцінювання, вибір найкращої або побудова узагальненої моделі ряду.

На основі побудованої моделі розраховують точкові та інтервальні прогнози.

Вирізняють **короткотермінове**, **середньотермінове** та **довготермінове** прогнозування. Прикладом короткотермінового прогнозу є розрахунок основних макроекономічних показників на період півтора роки, що здійснюється у середині кожного року з метою формування державного бюджету наступного року. Зазвичай в основу такого прогнозування покладають екстраполяцію наявних тенденцій динаміки відповідного показника.

Прикладом завдання довготермінового прогнозування є розробка великого інвестиційного проекту (міжнародна космічна станція, будівництво експериментального термоядерного реактора тощо). У таких ви-

падках необхідно попередньо здійснити аналіз динаміки окремих істотних для реалізації проекту складових – рівня інфляції, курсів валют, цін на основні ресурси й т.п. Зазвичай при довготерміновому прогнозуванні необхідно здійснити декомпозицію досліджуваних часових рядів і побудувати їх адекватні моделі. Інтервали часу, характерні для різних типів прогнозів, істотно залежать від довжини наявних рядів, а також особливостей динаміки аналізованих процесів.

Розробка прогнозу потребує оцінювання його точності й надійності. Найпростішою мірою точності прогнозу є його абсолютна похибка, тобто різниця між прогнозованим та фактичним значеннями досліджуваного показника. Але це можливо лише за умови, що період прогнозування вже пройшов, або у випадку, коли прогноз робиться для певного моменту часу в минулому. Перший варіант є непридатним, якщо результати прогнозування необхідно використовувати для підготовки відповідальних рішень, які мають бути прийнятими до того, як з'явиться можливість перевірки якості прогнозу. Другий варіант використовується переважно для попередньої перевірки адекватності застосовуваних моделей досліджуваних рядів.

Основними мірами точності прогнозу є:

– його середньоквадратична похибка за n рівнів:

$$MSE = \frac{1}{n} \sum_{t=T+1}^{T+n} (y_t - \hat{y}_t)^2 ; \qquad (7.1)$$

– квадратний корінь із його середньоквадратичної похибки за n рівнів:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=T+1}^{T+n} (y_t - \hat{y}_t)^2} ; \qquad (7.2)$$

– середня абсолютна похибка за n рівнів:

$$MAD = \frac{1}{n} \sum_{t=T+1}^{T+n} \left| y_t - \hat{y}_t \right| ; \qquad (7.3)$$

– квадратний корінь із його середньоквадратичної похибки, взятої у відсотках від фактичних значень, за n рівнів:

$$RMSPE = 100\sqrt{\frac{1}{n}\sum_{t=T+1}^{T+n}\left(\frac{y_t - \hat{y}_t}{y_t}\right)^2}; \qquad (7.4)$$

– середня абсолютна похибка, взята у відсотках від фактичних значень, за n рівнів:

$$MAPE = \frac{100}{n}\sum_{t=T+1}^{T+n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|. \qquad (7.5)$$

На основі двох останніх величин розроблено градацію якості моделей (табл. 7.1).

Таблиця 7.1

Класифікація прогнозів за якістю

MAPE, RMSPE	Точність прогнозу
< 10%	Висока
10 – 20%	Добра
20 – 40%	Задовільна
40 – 50%	Погана
> 50%	Незадовільна

Якість прогнозу можна охарактеризувати також за допомогою його довірчої імовірності, що характеризує ступінь впевненості у тому, що досліджувана величина потрапить до певного інтервалу. Вона змінюється від нуля до одиниці (від нуля до 100 %). Збільшення довірчої імовірності призводить до розширення інтервального прогнозу і, відповідно, до зменшення його корисності. З іншого боку, при надмірно малих значеннях довірчої імовірності корисність прогнозу також є низькою. Тому можна говорити про певну оптимальну величину цього показника, яка залежить від конкретного завдання дослідження. Іншою відмінною рисою цього показника є те, що його можна застосовувати лише після побудови адекватної математичної моделі досліджуваного процесу.

Отримати статистичні оцінки точності прогнозів можна тільки при використанні **ретропрогнозу**, сутність якого полягає у тому, що прогноз будують за першими N – k рівнями ряду, а потім порівнюють прогнозовані значення для k рівнів, які залишилися, з фактичними.

При побудові точкових та інтервальних прогнозів на момент часу T + n розглядають два основні варіанти:

– випадкова компонента є білим шумом;
– випадкова компонента має внутрішню структуру, що описується ARMA моделлю.

У першому випадку точковий прогноз буде дорівнювати точковому прогнозу детермінованої компоненти ряду:

$$\hat{y}_{T+n} = \hat{u}_{T+n} + \hat{s}_{T+n} + \hat{v}_{T+n} = \hat{d}_{T+n}, \qquad (7.6)$$

а інтервальний будують на основі точкового за методикою регресійного аналізу.

У другому випадку точковий прогноз дорівнюватиме сумі точкових прогнозів детермінованої та випадкової компонент:

$$\hat{y}_{T+n} = \hat{u}_{T+n} + \hat{s}_{T+n} + \hat{v}_{T+n} + \hat{\varepsilon}_{T+n} = \hat{d}_{T+n} + \hat{\varepsilon}_{T+n}. \qquad (7.7)$$

Прогнозні значення випадкової компоненти для MA(q) процесу визначають за формулою:

$$\hat{\varepsilon}_{T+n} = \begin{cases} \mu + \theta_n \varepsilon_T + \theta_{n+1}\varepsilon_{T-1} + ... + \theta_{n+q}\varepsilon_{T-q} \quad \left(n = \overline{1,q}\right); \\ \mu \quad \left(n > q\right), \end{cases} \qquad (7.8)$$

а їх похибку – за формулою:

$$MSE = \begin{cases} \sigma^2 \quad \left(n = 1\right); \\ \sigma^2(1 + \theta_1^2 + \theta_2^2 + ... + \theta_{n-1}^2) \quad \left(n = \overline{2,q}\right); \\ \sigma^2(1 + \theta_1^2 + \theta_2^2 + ... + \theta_{q-1}^2) \quad \left(n > q\right). \end{cases} \qquad (7.9)$$

Для AR(1) процесу прогнозні значення та їх похибки розраховують за формулами:

$$\hat{u}_{T+n} = \mu + \varphi_1^n \varepsilon_T + \varphi_1^{n+1}\varepsilon_{T-1} + \varphi_1^{n+2}\varepsilon_{T-2} + ... \qquad 7.10)$$

$$MSE = \begin{cases} \sigma^2 \quad \left(n = 1\right); \\ \sigma^2(1 + \varphi_1^2 + \varphi_1^4 + ... + \varphi_1^{2(n-1)}) \quad \left(n > 1\right). \end{cases} \qquad (7.11)$$

Інтервальний прогноз з рівнем надійності $1-\alpha$ визначають за формулою:

$$\hat{y}_{T+n} - C_\alpha \sqrt{MSE} < y_{T+n} < \hat{y}_{T+n} + C_\alpha \sqrt{MSE}, \qquad (7.12)$$

де C_α коефіцієнти, наведені в табл. 7.2.

Таблиця 7.2

Значення C_α для побудови інтервальних прогнозів

$1-\alpha$	0,8	0,9	0,95	0,99	0,999
C_α	1,28	1,64	1,96	2,57	3,29

Визначити якість прогнозної моделі можна лише після закінчення прогнозного періоду, а обирати її зазвичай необхідно до його початку. Часто можна отримати декілька адекватних моделей часового ряду, що дають прогнози, які мало розрізняються за характеристиками точності. Для вирішення цієї проблеми застосовують дисперсійно-коваріаційний і регресійний методи побудови комбінованих (узагальнених) прогнозів, які дають змогу врахувати результати всіх побудованих прогнозів і отримати результат, який у певному розумінні буде кращим за всі інші прогнози.

Оптимальний прогноз будують як лінійну комбінацію наявних оцінок F_i: $F = \sum_{i=1}^{k} \lambda_i F_i$. При цьому ставиться завдання визначення вагових коефіцієнтів λ_i.

У дисперсійно-коваріаційному методі за наявності двох прогнозів їх розраховують згідно з формулами:

$$\lambda_1 = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}, \quad \lambda_2 = \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}, \qquad (7.13)$$

де $\sigma_1^2 = \frac{\sum u_{1t}^2}{T}$ та $\sigma_2^2 = \frac{\sum u_{2t}^2}{T}$ – дисперсії похибок прогнозів, $\sigma_{12} = \frac{\sum u_{1t} u_{2t}}{T}$ – коваріація цих похибок, $u_{1t} = y_t - F_{1t}$ та $u_{2t} = y_t - F_{2t}$ – похибки прогнозів.

Сутність регресійного методу оптимізації прогнозу полягає у визначенні параметрів регресійних моделей:

$$y_t = \lambda_1 F_{1t} + \lambda_2 F_{2t} + v_t \qquad (7.14)$$

або

$$y_t = \lambda_0 + \lambda_1 F_{1t} + \lambda_2 F_{2t} + v_t \qquad (7.15)$$

для незміщених і зміщених прогнозів, відповідно. Знайдені параметри λ_1, λ_2 потім використовують як вагові коефіцієнти. Для незміщених прогнозів вагові коефіцієнти, отримані цим методом, збігаються з розрахованими за формулами (7.13).

7.2. Адаптивне прогнозування [Горчаков А.А. Математический аппарат для инвестора // Аудит и финансовый анализ. – 1997. – № 3. – С. 1 - 57].

При короткотерміновому прогнозуванні зазвичай більш важливою є динаміка досліджуваного показника на кінці періоду спостережень. Тому у таких випадках часто використовують адаптивні методи, що враховують інформаційну нерівноцінність даних. Параметри базової адаптивної моделі визначають за декількома першими спостереженнями і роблять прогноз, який порівнюють з відомими значеннями показника. Далі модель корегують відповідно до відхилень прогнозних значень від фактичних і роблять прогноз на наступний період. Цю процедуру повторюють до повного вичерпання наявних даних. У цьому разі прогноз можна розглядати як екстраполяцію останньої тенденції у динаміці досліджуваного показника. Як базову адаптивну модель зазвичай обирають моделі Брауна, Хольта або авторегресійну. Перші дві моделі належать до ARMA моделей. Існує багато методик прогнозування, що побудовані на основі цих моделей і розрізняються способами оцінювання параметрів, алгоритмами адаптації та компоновкою.

У **моделі Брауна** прогноз у момент часу t на момент часу $t + \tau$ отримують за формулою:

$$\overline{x}_t(t+\tau) = a_{1,t} + a_{2,t}, \qquad (7.16)$$

де

$$a_{1,t} = a_{1,t-1} + a_{2,t-1} + (1-\beta^2)e_t;\ a_{2,t} = a_{2,t-1} + (1-\beta^2)e_t, \qquad (7.17)$$

β – коефіцієнт дисконтування даних, $e_t = x_t - \overline{\overline{x}}_{t-1}(t)$ - похибка прогнозу.

Початкові значення визначають методом найменших квадратів на основі декількох перших спостережень. Значення параметру дисконтування, що лежить в інтервалі від нуля до одиниці, вибирають шляхом чисельної мінімізації дисперсії похибок. При цьому воно має бути сталим для всього процесу адаптації.

Модель Брауна можна трактувати як АРМА модель з параметрами: p = 0, d = 2, q = 2.

В **моделі Хольта** коефіцієнти рівняння (7.16) визначають по формулам:

$$a_{1,t} = a_{1,t-1} + a_{2,t-1} + \gamma_1 e_t;\ a_{2,t} = a_{2,t-1} + \gamma_2 e_t, \qquad (7.18)$$

де $\gamma_1, \gamma_2 \in (0,1)$ - параметри згладжування. Їх оптимальні параметри, які є сталими для всього періоду спостережень, визначають методами багатовимірної чисельної оптимізації.

В **моделі авторегресії порядку p** значення поточного рівня ряду подають як зважену суму p попередніх спостережень:

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_p x_{t-p}. \qquad (7.19)$$

Параметри простої авторегресійної моделі оцінюють методом найменших квадратів, а для більш складних моделей використовують інші методи. Початкове значення порядку авторегресії визначають на основі аналізу автокореляційної функції, а потім уточнюють методом перебору варіантів. Найкращим вважають таке значення, що забезпечує мінімальну дисперсію похибок. У сезонних моделях порядок беруть рівним періоду сезонності.

У багатьох випадках сезонні моделі з оцінками параметрів по методу найменших квадратів є надмірно складними. Для підвищення стійкості моделі у багатьох випадках доцільно будувати її для рядів, що попередньо зведені до стаціонарності. Прикладом таких моделей є **модель Бокса-Дженкінса**.

7.3. Проблеми агрегування й дезагрегування часових рядів

Значна кількість часових рядів, що зустрічаються у практиці, подана у вигляді щоквартальних або щорічних даних. Іноді виникає проблема їх зіставлення з даними, що мають більш дрібну, наприклад, щомісячну, структуру. Нерідко також частина ряду подається у вигляді

щоквартальних даних, а його інша частина – у вигляді щорічних. Для розв'язання цих проблем необхідно застосовувати методи агрегування або дезагрегування часових рядів. У першому разі ряд з більш дрібною структурою приводять до вигляду, що має інший ряд, шляхом об'єднання даних, які відносяться до одного і того самого року. При цьому значна частина інформації, що міститься в першому ряді, втрачається.

Розглянемо задачу створення ряду $x_1, x_2, ..., x_{4T}$, що містить щоквартальні значення певного показника, на основі наявного ряду $y_1, y_2, ..., y_T$ з річною структурою даних, виходячи з умови $y_t = x_{4t-3} + x_{4t-2} + x_{4t-1} + x_{4t}$. Для її розв'язання застосовують методи процентного відношення і поліноміальної інтерполяції.

Нехай для певного року y_k є відомими квартальні значення $z_{k1}, z_{k2}, z_{k3}, z_{k4}$. Тоді, згідно з методом процентного відношення, елементи нового ряду можна отримати за формулою:

$$x_{4t-4+i} = y_t \cdot \alpha_i, \; i = \overline{1,4}, \qquad (7.20)$$

де $\alpha_i = \dfrac{z_{ki}}{y_k}$.

Така методика передбачає, що ряд має пропорційну структуру, тобто процентні співвідношення між значеннями показника у різних кварталах є сталими величинами.

Якщо є щоквартальні дані за декілька років, номери яких утворюють множину $S = \{t_1, t_2, ..., t_n\}$, то коефіцієнти α_i розраховують як середнє геометричне: $\alpha_i = \sqrt[n]{\prod_{k \in S} \dfrac{z_{ki}}{y_k}}$, $i = \overline{1,4}$. У разі, коли для жодного року немає щоквартальних даних, тобто множина S є пустою, припускають, що $\alpha_i = 1/4, \; i = \overline{1,4}$.

Метод поліноміальної інтерполяції передбачає створення на першому етапі допоміжного ряду $v_j = \sum_{t=1}^{j} y_t, \; j = \overline{1,T}$, що складається з акумулятивних сум вихідного ряду. Після цього за допомогою одного із стандартних методів будують інтерполяційну функцію f(t) і обчислюють її значення у проміжних точках, що відповідають окремим кварталам.

Практичне застосування розглянутих методів дає можливість зробити таки зауваження:

– метод процентного відношення може додавати сезонні коливання до рядів, які їх не мали; за відсутності поквартальних даних цей метод будує ряд, який має східчастий вигляд;

– метод поліноміальної інтерполяції здійснює згладжування вихідного ряду й може усувати сезонні коливання; він також може додавати відсутню у вихідному ряді слабку циклічну складову; поквартальні дані для цього методу не є необхідними, але їх можна використовувати для визначення коефіцієнта сезонності при побудові прогнозу; метод поліноміальної інтерполяції не дає можливості розбити на квартальні дані спостереження, що відповідають першому року.

У зв’язку із викладеним вище рекомендується застосовувати метод процентного відношення за наявності поквартальних даних, що мають сезонну компоненту, за декілька років. Якщо ж сезонних коливань немає, або є лише річні дані, доцільнішим буде використання методу поліноміальної інтерполяції.

7.4. Аналіз варіабельності кардіограм

Однією з основних ідей фізіології є спрямованість нормальної життєдіяльності організму на зменшення його змінюваності (варіабельності). Зокрема, для опису серцевої діяльності застосовують термін "регулярний синусовий ритм", який означає сталість проміжку часу між двома послідовними скороченнями серця. При деяких захворюваннях регулярність ритму порушується. На рис. 7.1 наведено часові ряди, що характеризують проміжки часу між двома послідовними скороченнями для здорового (ряд 1) і хворого (ряд 2) серця, побудовані за даними [23].

До стандартних статистичних характеристик динамічних рядів кардіоінтервалів належать: середнє значення (MV), стандартне відхилення кардіоінтервалів (SDNN), квадратний корінь із середнього квадрату різниці величин послідовних пар RR-інтервалів (RMSSD), відсоток кількості пар послідовних кардіоінтервалів, які відрізняються один від одного більше ніж на 50 мс (PNN50), а також коефіцієнт варіації (CV). При цьому із статистичного аналізу виключають випадкові передчасні скорочення (екстрасистоли). У нормі їх значення знаходяться у межах: SDNN – 30–100 мс, RMSSD – 20–50 мс, CV – 3–12 %.

Відповідні дані для динамічних рядів, показаних на рис. 7.1, наведені у табл. 7.3.

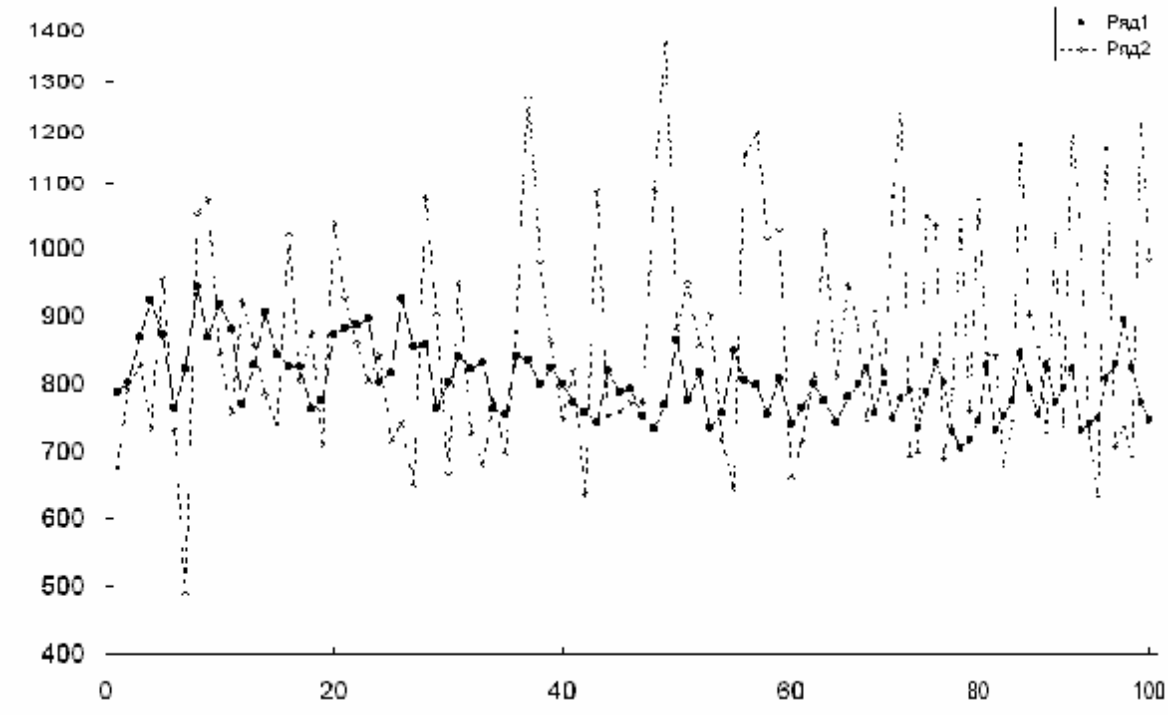

Рис. 7.1. Динамічні ряди інтервалів між послідовними скороченнями серця

Таблиця 7.3

Статистичні характеристики динамічних рядів, наведених на рис. 7.1

	MV, мкс	SDNN, мкс	RMSSD, мкс	PNN50, %	CV, %
Ряд 1	803	52	54	43,4	
Ряд 2	869	172	239	81,8	

Результати аналізу зручно подавати у вигляді гістограми, до якої додають моду розподілу Mo і варіаційний розкид MxDMn. Для обчислення моди спочатку визначають модальний інтервал гістограми, тобто інтервал, для якого частота має найбільше значення. Потім моду розраховують за формулою:

$$Mo = x_0 + \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})}\Delta \ , \qquad (7.21)$$

де x_0 – нижня межа модального інтервалу; f_{Mo} – частота в модальному інтервалі; f_{Mo-1} – частота в попередньому інтервалі; f_{Mo+1} – частота в інтервалі, що йде за модальним; Δ – величина інтервалу.

Для візуального порівняння різних гістограм вихідні дані групують за 20 інтервалами рівної величини, що охоплюють діапазон від 400 до 1300 мс. За даними аналізу гістограми обчислюють стрес-індекс (відношення висоти гістограми до її ширини), що характеризує ступінь напруження регуляторних систем організму. У нормі його величина знаходиться у межах 50–150 ум.од.

Як видно з даних, наведених на рис. 7.1, 7.2 та в табл. 7.3, патологія, що розглядається, призводить до істотного збільшення середнього значення й дисперсії часового ряду. У загальному випадку необхідно перевірити нульову гіпотезу, що досліджувані ряди є двома випадковими реалізаціями одного й того самого процесу, тобто є вибірками з однієї й тієї самої генеральної сукупності.

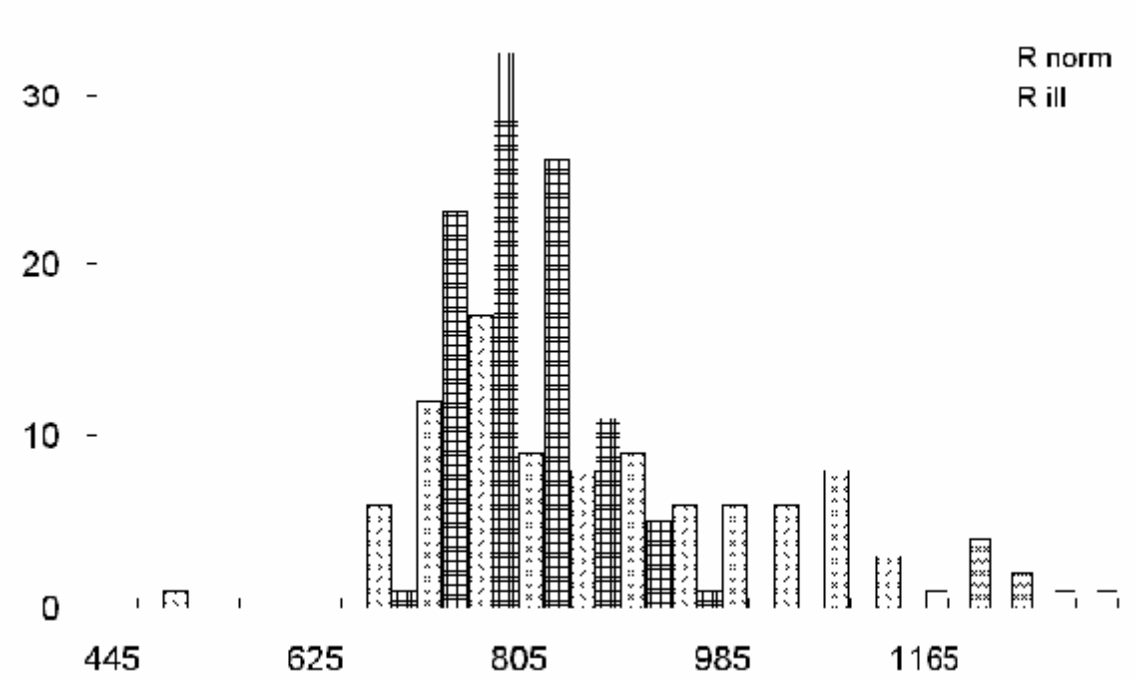

Рис. 7.2. Гістограми динамічних рядів інтервалів між послідовними скороченнями для здорового (R_{norm}) та хворого (R_{ill}) серця