# Міністерство освіти і науки України КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

# МЕТОДИЧНІ РЕКОМЕНДАЦІЇ

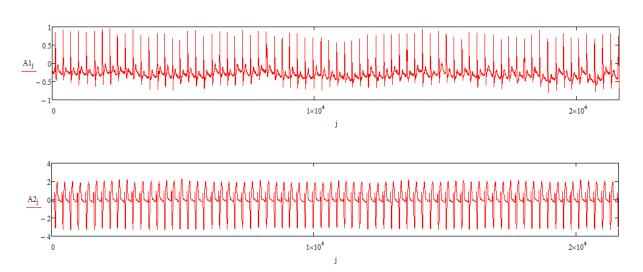
до виконання розрахункової роботи з дисципліни "Інтелектуальна обробка даних"

професор кафедри теоретичної кібернетики, д.ф.-м.н., професор Пашко Анатолій Олексійович

**Завдання.** У файлах A1.txt- A32.txt міститься запис кардіограми людини по 12 каналах. Час запису — 10 сек. Дискретність: 500 точок за 1 сек. Структура файлу — 1-й канал, 2-й канал, ... 12-й канал (амплітуда у відносних одиницях). Довжина запису N=5000,  $\Delta t = 1/500 = 0.002$ .

### Алгоритм обробки

1. Побудувати графік кардіограми по кожному каналу. Для заданих змінних оцінити основні статистичні параметри (середнє, середнє гармонічне, середнє геометричне, дисперсію, середню різницю Джині, моду, медіану, коефіцієнт асиметрії, коефіцієнт ексцесу, побудувати гістограму, перевірити гіпотезу про закон нормальний закон розподілу). Нормалізувати дані по кожному стовпчику (математичне сподівання рівне нулю, дисперсія рівна 1).



Провести візуальне порівняння.

## 2. Однофакторний дисперсійний аналіз.

Перевірити чи  $\epsilon$  результати вимірювання різними рівнями одного фактору (12 рівнів).

Для кожного рівня знаходимо

$$S_i^2 = \frac{1}{n-1} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 = \frac{1}{n-1} \left[ \sum_{j=1}^n x_{ij}^2 - \frac{1}{n} \left( \sum_{j=1}^n x_{ij} \right)^2 \right].$$

За припущенням дисперсійного аналізу - повинна мати місце рівність дисперсій. Перевірити рівність дисперсій.

При виконанні припущення про рівність дисперсій, знаходимо оцінку дисперсії, що характеризує розсіювання поза впливом фактора,

$$S_0^2 = \frac{1}{k} \sum_{i=1}^k S_i^2 \frac{1}{k(n-1)} \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 = \frac{1}{k(n-1)} \left[ \sum_{i=1}^k \sum_{j=1}^n x_{ij}^2 - \frac{1}{n} \sum_{i=1}^k \left( \sum_{j=1}^n x_{ij} \right)^2 \right]$$

Знаходимо вибіркову дисперсію всіх спостережень

$$S^{2} = \frac{1}{kn-1} \left[ \sum_{i=1}^{k} \sum_{j=1}^{k} x_{ij}^{2} - \frac{1}{kn} \left( \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^{2} \right]$$

Знаходимо оцінку дисперсії, що характеризує зміни параметра, пов'язані з фактором

$$S_A^2 = \frac{n}{k-1} \sum_{i=1}^k (\bar{x}_i - \bar{\bar{x}})^2 . \qquad \bar{\bar{x}} = \frac{1}{k} \sum_{i=1}^k \bar{x}_i; \quad \bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij}.$$

Оцінка впливу фактора на зміни середнього значення визначається відношенням (вплив значущий з ймовірністю 1-а)

$$\frac{S_A^2}{S_0^2} > F_{\alpha}[k-1; k(n-1)]$$

де  $F_{\alpha}(f_1, f_2)$  -  $\alpha$ - квантиль F- розподілу з  $f_1$  та  $f_2$  степенями свободи.

#### 3. Двохфакторний аналіз.

Побудувати таблицю двохфакторного експерименту за правилом – кожен канал розбити на 5 частин (по 1000 даних у кожній частині)

Рівні	Рівні фактора А (і)								
фактора	$A_1$	$A_2$	$A_3$	A <sub>4</sub>	$A_{12}$				
фактора В(j)									
$B_1$									
$B_2$									
$B_3$									
$B_4$									
$B_5$									

В кожній клітинці таблиці зберігається масив із n=1000 значень.

Якщо фактори А і В незалежні:

Знаходимо середнє значення в кожній клітинці хіі.

Обчислюємо основні показники

$$Q_{1} = \sum_{i=1}^{k} \sum_{j=1}^{m} x_{ij}^{2}; \quad Q_{2} = \frac{1}{m} \sum_{i=1}^{k} X_{i}^{2}; \quad Q_{3} = \frac{1}{k} \sum_{j=1}^{m} X_{j'}^{2};$$
$$Q_{4} = \frac{1}{mk} \left( \sum_{i=1}^{k} X_{i} \right)^{2} = \frac{1}{mk} \left( \sum_{j=1}^{m} X_{j'} \right)^{2}.$$

 $X_i$  - сума по стовпчиках (рівень фактора  $A_i$ ),  $X_{i'}$  - сума по рядках (рівень фактора  $B_i$ ).

Знаходимо оцінки дисперсій

$$S_0^2 = \frac{Q_1 + Q_4 - Q_2 - Q_3}{(k-1)(m-1)}; \quad S_A^2 = \frac{Q_2 - Q_4}{k-1}; \quad S_B^2 = \frac{Q_3 - Q_4}{m-1}.$$

Якщо  $\frac{S_A^2}{S_0^2} > F_{\alpha}(f_1,f_2)$  для  $f_1$ =k-1,  $f_2$ =(k-1)(m-1) то фактор A  $\varepsilon$  значущим. Аналогічно для фактора B при  $f_1$ =m-1,  $f_2$ =(k-1)(m-1) .

Якщо фактори А і В залежні:

Знаходимо додатково

$$Q_5 = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{\nu=1}^{n} x_{ij\nu}^2$$
.  $S_{AB}^2 = \frac{Q_5 - nQ_1}{mk(n-1)}$ ,

взаємодію факторів перевіряємо за критерієм

$$\frac{nS_0^2}{S_{AB}^2} > F_{lpha}(f_1, f_2)$$
 , де  $f_1 = (k-1)(m-1)$  ,  $f_2 = mk(n-1)$  .

# 4. Перетворення Фур'є.

Виконати перетворення Фур'є за формулами

$$A_{\delta} := \frac{1}{N} \cdot \sum_{i=0}^{N-1} \left( s1_i \cdot \cos\left(\frac{2\pi \cdot i \cdot 0}{N}\right) \right) \qquad A_{\frac{N}{2}} := \frac{1}{N} \cdot \sum_{i=0}^{N-1} \left( s1_i \cdot \cos\left(\frac{\pi \cdot i}{1}\right) \right)$$

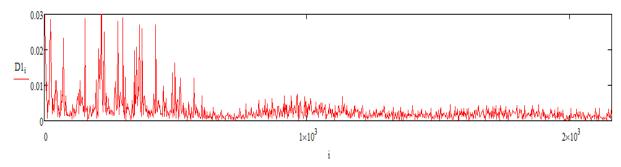
$$A_1 := \frac{2}{N} \cdot \sum_{i=0}^{N-1} \left( s1_i \cdot cos\left(\frac{2\pi \cdot i \cdot 1}{N}\right) \right)$$

$$1 := 1, 2... \frac{N}{2} - 1$$

$$B_{j} := \frac{2}{N} \cdot \sum_{i = 0}^{N-1} \left( s1_{i} \cdot sin \left( \frac{2\pi \cdot i \cdot j}{N} \right) \right) \\ j := 0, 1... \frac{N}{2}$$

$$C_{\hat{j}} := \sqrt{\left(A_{\hat{j}}\right)^2 + \left(B_{\hat{j}}\right)^2} \qquad j := 0, 1... \frac{N}{2}$$

Знайти спектр



Побудувати графік для кожної змінної. Обчислити частоту першої синусоїди (або крок по частоті).

Порівняти отримані результати.

Виконати обернене перетворення Фур'є

$$d1_{i} := \sum_{j=0}^{\frac{N}{2}} \left( A_{j} \cdot cos\left(\frac{2\pi j \cdot i}{N}\right) \right) + \sum_{j=0}^{\frac{N}{2}} \left( B_{j} \cdot sin\left(\frac{2\pi j \cdot i}{N}\right) \right)$$

Порівняти початковий і результуючий масиви.

### 5. Кореляційний аналіз

Крок 1.. Нормалізація всіх змінних:

$$x_{ij}^* = \frac{x_{ij} - \overline{x}_j}{\sigma_{x_j}}, i = \overline{1, N}; j = \overline{1, m}.$$

Крок2. Обчислення кореляційної матриці

$$r = \frac{1}{n} (X^* X^*).$$

		0	1	2	3	4	5	6	7	8	9	10	11
	0	1	0.915	0.269	-0.972	0.88	0.802	-0.461	0.038	0.755	0.907	0.917	0.924
	1	0.915	1	0.617	-0.983	0.628	0.972	-0.449	-0.072	0.671	0.881	0.9	0.912
	2	0.269	0.617	1	-0.473	-0.201	0.773	-0.167	-0.223	0.157	0.356	0.377	0.393
	3	-0.972	-0.983	-0.473	1	-0.752	-0.916	0.462	0.025	-0.725	-0.914	-0.929	-0.938
	4	0.88	0.628	-0.201	-0.752	1	0.444	-0.388	0.15	0.687	0.741	0.744	0.744
dk =	5	0.802	0.972	0.773	-0.916	0.444	1	-0.416	-0.131	0.569	0.797	0.82	0.835
	6	-0.461	-0.449	-0.167	0.462	-0.388	-0.416	1	0.827	0.157	-0.326	-0.446	-0.506
	7	0.038	-0.072	-0.223	0.025	0.15	-0.131	0.827	1	0.594	0.103	-0.029	-0.098
	8	0.755	0.671	0.157	-0.725	0.687	0.569	0.157	0.594	1	0.86	0.785	0.738
	9	0.907	0.881	0.356	-0.914	0.741	0.797	-0.326	0.103	0.86	1	0.99	0.975
	10	0.917	0.9	0.377	-0.929	0.744	0.82	-0.446	-0.029	0.785	0.99	1	0.995
	11	0.924	0.912	0.393	-0.938	0.744	0.835	-0.506	-0.098	0.738	0.975	0.995	1

Рис. 1 Приклад кореляційної матриці

**Крок 3.** Аналіз кореляційної матриці — виділити групу із трьох (чотирьох) параметрів, парна кореляція між якими велика (коефіцієнт кореляції близька по модулю до 1).

Наприклад, параметри 0, 1, 3 або 0, 1, 3, 11 (рис. 1).

Параметри a, b, c та d вибирати в залежності від результатів розрахунків по варіантах.

**Крок 4.** Знайти частковий коефіцієнт кореляції між ознаками a та b без урахування впливу ознаки c:

$$r_{ab(c)} = \frac{r_{ab} - r_{ac}r_{bc}}{\sqrt{(1 - r_{ac})(1 - r_{bc})}}$$

 $r_{ij}$  - коефіцієнт кореляції (кореляція Пірсона) між параметрами i та j.

Крок 5. Знайти частковий коефіцієнт кореляції:

- між ознаками a та c без урахування впливу ознаки b;
- між ознаками a та b без урахування впливу ознаки c та d;
- між ознаками a та c без урахування впливу ознаки b та d;
- між ознаками a та d без урахування впливу ознаки b та c.

**Крок 6.** Знайти множинний коефіцієнт кореляції для параметра a, при лінійному двофакторному зв'язку з параметрами b та c

$$r_{a/bc} = \sqrt{\frac{r_{ab}^2 + r_{ac}^2 - 2r_{ab}r_{ac}r_{bc}}{1 - r_{bc}^2}}$$
.

**Крок 7.** Знайти множинний коефіцієнт кореляції для параметра a, при лінійному трифакторному зв'язку з параметрами b, c та d.

$$R_{a/bcd} = 1 - (1 - r_{ab}^2)(1 - r_{ac(b)}^2)(1 - r_{ad(bc)}^2)$$

**Крок 8.** Зробити висновки про кореляцію між параметрами. Чи існує параметр, що  $\epsilon$  незалежним (кореляція з іншими факторами невелика)?

# 6. Факторний аналіз

**Крок 1.** Знаходження власних чисел кореляційної матриці r з рівняння

$$|r-\lambda E|=0, k=\overline{1,m},$$

де E — одинична матриця розміром  $m \times m$ .

Власні числа та частка дисперсії, що пояснюється змінною

№п/п	Власні числа	Частка дисперсії	Сумарна дисперсія
1	0.00123	11.9997	100
2	0.001335	11.9985	99.98974975
3	0.0004547	11.9971	99.97862449
4	0.0005151	11.9967	99.97483523
5	0.002178	11.9962	99.97054263
6	0.011	11.994	99.95239219

7	0.016	11.983	99.86072333
8	0.027	11.967	99.72738681
9	0.269	11.94	99.50238142
10	1.386	11.671	97.26066111
11	2.23	10.285	85.71038467
12	8.055	8.055	67.12660656

Перші три змінні пояснюють 97% всієї дисперсії.

**Крок 2.** Власні значення  $\lambda_k$  упорядковуються за абсолютним рівнем вкладу кожного головного компонента до загальної дисперсії.

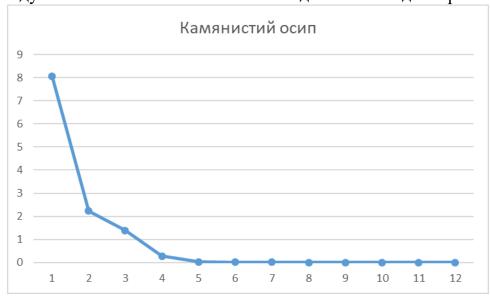


Рис. Графік для критерію кам'янистого осипу

Критерій інформативності приймає значення

$$I_k = \frac{\lambda_1' + \lambda_2' + \lambda_3'}{\lambda_1' + \lambda_2' + \lambda_3' + \dots + \lambda_{10}' + \lambda_{11}' + \lambda_{12}'} = \frac{11,671}{12,0} = 0,9726.$$

**Крок 3.** Обчислення власних векторів  $a_k$  розв'язуємо систему рівнянь  $(r - \lambda E)a = 0$ 

за таких умов:

$$a_j'a_k = \begin{cases} 0 (j \neq k), \\ 1 (j = k). \end{cases}$$

### Власні вектори (матриця L)

		0	1	2	3	4	5	6	7	8	9	10	11
	0	-0.607	0.076	0.174	5.393·10 ·3	-0.1	0.533	0.237	-0.074	-0.316	-0.149	0.048	-0.34
	1	0.386	0.027	0.674	0.152	0.023	0.018	-0.342	-0.206	-0.262	0.152	-0.074	-0.34
	2	0.149	-0.023	0.014	-1.553·10 <sup>-3</sup>	0.075	0.068	0.558	0.317	-0.041	0.678	-0.264	-0.158
	3	-0.404	0.033	0.645	0.161	-0.011	-0.349	0.176	0.204	0.278	-0.021	0.02	0.348
	4	0.214	-0.019	6.152·10 ·3	0.025	0.22	-0.453	0.507	0.038	-0.335	-0.476	0.175	-0.268
vv =	5	-0.481	-0.014	-0.218	-0.025	0.017	-0.589	-0.308	-0.068	-0.241	0.309	-0.142	-0.315
	6	-0.03	0.046	2.18-10-3	-0.024	0.084	-0.057	0.254	-0.704	0.035	0.365	0.516	0.16
	7	0.076	0.359	-0.022	-9.957·10 <sup>-3</sup>	-0.326	-0.027	-0.148	0.484	-0.226	0.17	0.649	5.195·10 ·3
	8	-0.065	-0.758	0.056	-0.037	0.218	0.08	-0.118	0.218	0.226	0.062	0.416	-0.271
	9	-0.043	0.428	-0.139	0.533	0.439	0.07	-0.063	0.046	0.428	-0.011	0.1	-0.339
	10	0.012	0.296	0.157	-0.768	0.035	-0.062	0.013	-0.013	0.414	-0.049	0.015	-0.343
	11	0.084	-0.123	-0.061	0.271	-0.762	-0.135	0.172	-0.155	0.353	-0.063	-0.031	-0.345

Рис. Матриця власних векторів

_	
	0
0	0.34
1	0.34
2	0.158
3	-0.348
4	0.268
5	0.315
6	-0.16
7	-5.195·10 <sup>-3</sup>
8	0.271
9	0.339
10	0.343
11	0.345

Рис. Власний вектор максимального власного числа

**Крок 4.** Перевірити виконання умов  $a_j'a_k = \begin{cases} 0 \ (j \neq k), \\ 1 \ (j = k). \end{cases}$  для векторів  $a_k$ .

**Крок 5.** Знаходження головних компонентів (векторів  $z_k = x \cdot a_k$ ,  $k = \overline{1, m}$ .). Побудова гафіків основних компонент, перевірка властивостей. Перші три головні фактори

ricpi	перші три головні фактори								
№п/п	Z1	Z1 Z2							
1	0,143742	0,106662	-0,35059						
2	-0,48503	0,044493	-0,37265						
3	-0,31641	0,044899	-0,38265						

4	-0,25538	0,008308	-0,3653
5	-0,1269	0,04438	-0,37617
6	-0,30143	0,060948	-0,35856
7	0,133987	0,093535	-0,32858
8	0,312733	0,121627	-0,3473
• • •	•••	•••	•••
5000	6,240308	2,605887	4,476333

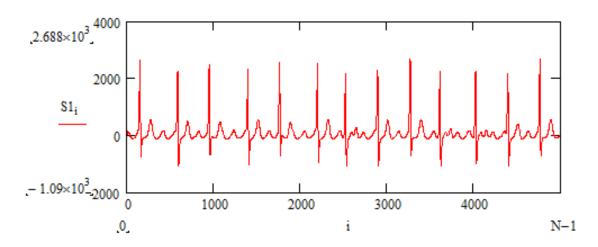


Рис. Графік першої головної компоненти

**Крок 6.** Головні компоненти задовольняють умови:

$$\sum_{i=1}^{n} z_{k,i} = 0, i = \overline{1,n};$$

$$\frac{1}{n} z'_{k} z_{k} = \lambda_{k}, k = \overline{1,m};$$

$$z'_{j} z_{k} = 0, j = \overline{1,m}, j \neq k.$$

Перевірити виконання вказаних умов.

Крок 7. Зробити висновки за результатами факторного аналізу.

## 7. Кластерний аналіз

А). Будемо вважати, що записана кардіограма (12 каналів) являє собою множину багатовимірних точок деякого евклідового простору. Розмірність точок -12, кількість точок -5000.

Використовуючи алгоритми кластеризації, а саме k-means, провести розбиття множини точок на k підмножин. Розглянути варіанти k=5, k=7.

Б) Провести кластеризацію даних трьох головних факторів (дані  $\epsilon$  трьох вимірними векторами)

Результатом  $\epsilon$  перелік точок, що входять до кожного кластеру. Визначити, чи співпадають кластери у варіантах A) та Б).

### Алгоритм k-Means (Hard-c-means)

- 1. Випадково обрати k точок, які будуть початковими центрами мас кластерів.
- 2. Віднести кожен об'єкт до кластеру з найближчим центром мас.
- 3. Перерахувати центри мас кластерів згідно з поточним членством.
- 4. Якщо критерій зупинки алгоритму не виконується, повернутися до кроку 2.

У якості критерію зупинки обирають один з двох варіантів:

- 1. Немає переходу об'єктів з кластера в кластер на кроці 2.
- 2. Мінімальна зміна середньоквадратичної помилки.

Алгоритм чутливий до вибору центрів мас.

### Рекомендована література Базова література

- 1. Джессен Р.Д. Методы статистических обследований / Р.Д. Джессен. М.: Финансы и статистика. 1985.
- 2. Бахрушин В.Є. Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. Запоріжжя : КПУ, 2011. 268 с.
- 3. Айвазян С. А. Прикладная статистика: Исследование зависимостей: Справ. изд. / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. М.: Финансы и статистика. 1985. 487 с.
- 4. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С. А. Айвазян, В.М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. М.: Финансы и статистика. 1989. 607 с.
- 5. Добеши И. Десять лекций по вейвлетам / И. Добеши. Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика». – 2004. – 464 с.
- 6. Бендат Дж. Прикладной анализ случайных данных / Дж. Бендат, А. Пирсол. М.: Мир. 1989. 540 с.
- 7. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. М.: ФИЗМАТЛИТ. 2006. 816 с.
- 8. Джонсон Н. Статистика и планирование эксперимента в технике и науке. Методы обработки данных / Н. Джонсон, Ф. Аннон. М.: Мир. 1980. 610 с.
- 9. Малла С. Вейвлеты в обработке сигналов / С. Малла. М.: Мир.- 2005.-672 с.
- 10. Майборода Р.Є. Регресія: Лінійні моделі: Навчальний посібник / Р.Є. Майборода. К.:ВПЦ «Київський університет». 2007. 296 с.
- 11. Ugarte M.D. Probability and statistics with R / M.D. Ugarte, A.F. Militino, A.T. Arnholt. Boca Raton, London, New York: CRC Press, Taylor&Francis Group. 2008. 700 p.

# Допоміжна література

- 12. Гнеденко Б.В. Курс теории вероятностей / Б.В. Гнеденко. М.: Наука. 1988.
- 13. Закс Л. Статистическое оценивание / Л. Закс. –М.: СТАТИСТИКА. 1976. 598 с.
- 14. Шеффе Г. Дисперсионный анализ / Г. Шеффе. М.: Наука. 1980. 512 с.
- 15. Закс Ш. Теория статистических выводов / Ш. Закс. М.: Мир. 1975. 776 с.
- 16. Кокрен У. Методы выборочного исследования / У. Кокрен. М.: Финансы и статистика. 1976.
- 17. Блаттер К. Вейвлет анализ. Основы теории / К. Блаттер. М.: Техносфера. 2004. 276 с.

Таблиця критичних значень критерію Фішера.

 $\alpha = 0.05$ 

lO	12	15	20	24	30	40	60	120	∞
242	244	246	248	249	250	251	252	253	254
19,4	19,4	19,4	19,4	19,5	1 <b>9,5</b>	19,5	19,5	19,5	19,5
8,79	8,74	8,70	8,66	8,64	<b>8,62</b>	8,59	8,57	8,55	8,53
5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
4,74	4,68	4,62	4,56	4,53	<b>4,</b> 50	4,46	4,43	4,40	4,36
4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00