МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет "Львівська політехніка"

Інтелектуальний аналіз даних за допомогою табличного процесора Excel. Кореляційний аналіз.

МЕТОДИЧНІ ВКАЗІВКИ
до лабораторної роботи № 8

з курсу **"Системи інтелектуального аналізу та візуалізації даних"**

для студентів за освітньою програмою Комп'ютерні науки (Проектування і програмування інтелектуальних систем та пристроїв)

Затверджено на засіданні кафедри
"Системи автоматизованого проектування"

Протокол N 1 від 28.08.2023р.

ЛЬВІВ 2023

1. МЕТА РОБОТИ

Мета роботи - повторити основні прийоми роботи з функціями Excel, навчитися обчислювати коефіцієнт кореляції, перевіряти його значимість і надійність за допомогою функцій Excel, навчитися обчислювати параметри прямого і оберненого прогнозів за допомогою функцій Excel, навчитися будувати графіки прогнозів за допомогою діаграм Excel, визначати кут між ними.

2. КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

2.1. Види і форми взаємозв'язку між явищами

Кореляція — це статистичний спосіб розгляду зв’язку. Коли два об'єкти співвідносяться, це означає, що вони змінюються разом.

Позитивна кореляція означає, що високі бали одного об'єкту асоціюються з високими балами іншого об'єкту. Діаграма розсіювання на рис.1 є прикладом позитивної кореляції.

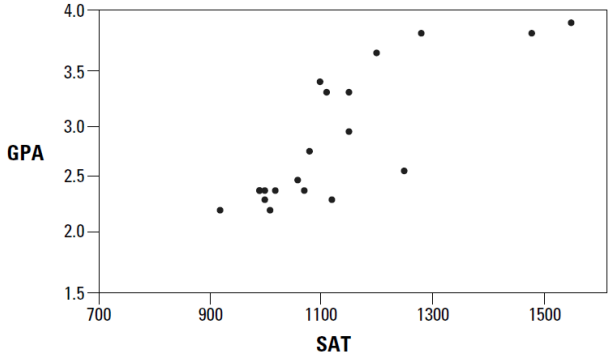


Рис.1. Діаграма розсіювання

З іншого боку, *негативна кореляція* означає, що високі бали одного об'єкту пов’язані з низькими балами іншого об'єкту. Прикладом є кореляція між масою тіла та часом, витраченим на програму схуднення. Якщо програма ефективна, то чим більше часу витрачено на програму, тим менше маса тіла. Крім того, чим менша кількість часу, витрачена на програму, тим більша вага тіла. Рис.2 показує дані з діаграми розсіювання на рис.1.

Student	*SAT*	*GPA*
1	990	2.2
2	1150	3.2
3	1080	2.6
4	1100	3.3
5	1280	3.8
6	990	2.2
7	1110	3.2
8	920	2.0
9	1000	2.2
10	1200	3.6
11	1000	2.1
12	1150	2.8
13	1070	2.2
14	1120	2.1
15	1250	2.4
16	1020	2.2
17	1060	2.3
18	1550	3.9
19	1480	3.8
20	1010	2.0
Mean	1126.5	2.705
Variance	26171.32	0.46
Standard Deviation	161.78	0.82

Рис.2. Набір даних успішності студентів

Інформація про те, що таке SAT і GPA і про зв'язок цих двох величин: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3090148/

Формула для обчислення кореляції між SAT і GPA (як і будь якими іншими об'єктами) є наступною:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r = \frac{\left[\frac{1}{N-1}\right]\sum(x-\bar{x})(y-\bar{y})}{s_x s_y}$$

Терм ліворуч, *r*, називається коефіцієнтом кореляції. Його також називають коефіцієнтом кореляції Пірсона на честь його творця Карла Пірсона.

Два терми в знаменнику праворуч — це стандартне відхилення змінної *x* і стандартне відхилення змінної *y*. Терм у чисельнику називається *коваріацією*. Інший спосіб записати цю формулу є наступний:

$$r = \frac{\text{cov}(x,y)}{s_x s_y}$$

Коваріація показує, як *x* і *y* змінюються разом. Ділення коваріації на добуток двох стандартних відхилень накладає певні обмеження. Нижня межа коефіцієнта кореляції –1,00, верхня – +1,00.

Коефіцієнт кореляції –1,00 представляє ідеальну негативну кореляцію (низькі *x*-показники, пов'язані з високими *y*-показниками, і високі *x*-показники, пов'язані з низькими *y*-показниками). Кореляція +1,00 представляє ідеальну позитивну кореляцію (низькі *x*-показники, пов'язані з низькими *y*-показниками, високі *x*-показники, пов'язані з високими *y*-показниками.) Кореляція 0,00 означає, що дві змінні не пов'язані.

Застосовуючи формулу до даних з рис.2 отримаємо наступний результат:

$$r = \frac{\left[\frac{1}{N-1}\right]\sum(x-\bar{x})(y-\bar{y})}{s_x s_y} =$$

$$\frac{\left[\frac{1}{20-1}\right]\ (990-1126.5)(2.2-2.705)+\ldots+(1010-1126.5)(2.0-2.705)}{(161.78)(0.82)} = .817$$

Давайте розбираємося, що означає отримане число? Приймемо до відома, що кореляція тісно пов'язана з регресією.

На рис.3 показана діаграма розсіювання з лінією, яка «найкраще відповідає» точкам. Нагадаємо, що через ці точки можна провести нескінченну кількість ліній. Який з них найкраща?

Щоб бути «найкращою», лінія має відповідати певному стандарту: якщо ми накреслимо відстані у вертикальному напрямку між точками та лінією, і піднесемо ці відстані до квадрату, а потім просумуємо отримані значення, то найкраще підходитиме та лінія, яка робить суму цих квадратів відстаней якомога меншою. Ця лінія називається лінією регресії.

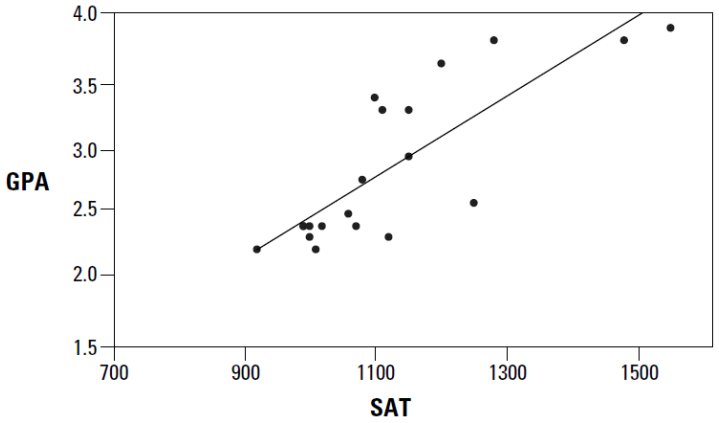


Рис.3 Діаграма розсіювання для оцінок 20 студентів, включаючи лінію регресії.

Нагадаємо, що мета лінії регресії в житті полягає в тому, щоб ми могли робити прогнози. Без лінії регресії нашим найкращим прогнозованим значенням змінної *y* є середнє значення *y*. Лінія регресії враховує змінну *x* і забезпечує точніший прогноз. Кожна точка на лінії регресії представляє прогнозоване значення для *y*. У символіці регресії кожне прогнозоване значення є *y*'.

Нагадаємо, теж, що у статистиці дисперсія вимірює відхилення від середнього значення. Щоб обчислити дисперсію необхіно виконати наступні дії:

- Обчислити середнє значення даних.
- Знайти різницю кожної точки даних від середнього значення.
- Піднести кожне з цих значень у квадрат.
- Просумувати усі значення, піднесені до квадрат.
- Розділити цю суму квадратів на *N* – 1 (для вибірки) або *N* (для сукупності).

Дисперсія обчислюється за такою формулою:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{N}$$

де:

x_i =Кожне значення в наборі даних

$\overline{x}$ = Середнє значення всіх значень у наборі даних

N=Кількість значень у наборі даних

Наприклад, під час обчислення дисперсії вибірки для оцінки дисперсії популяції (сукупності) знаменник рівняння дисперсії стає *N* – 1 для того, щоб оцінка була неупередженою та не недооцінювала дисперсію генеральної сукупності.

Для чого використовується дисперсія? Дисперсія – це, по суті, ступінь розкиду у наборі даних щодо середнього значення цих даних. Він показує кількість варіацій, які існують між точками даних. Візуально, чим більша дисперсія, тим «жирнішим» буде розподіл ймовірностей. Наприклад у бізнесі, якщо якась операція щось на кшталт інвестицій має значну дисперсію, це може бути інтерпретовано як ризикована або нестійка операція.

Стандартне відхилення використовується частіше ніж дисперсія. Чому?

Стандартне відхилення – це квадратний корінь із дисперсії. Іноді це корисніше, оскільки обчислення квадратного кореня видаляє одиниці з аналізу. Це дозволяє проводити прямі порівняння між різними величинами, які можуть мати різні одиниці виміру або різні величини.

Наприклад, сказати, що збільшення *X* на одну одиницю збільшує *Y* на два стандартні відхилення, дає змогу зрозуміти зв'язок між *X* і *Y* незалежно від того, в яких одиницях вони виражені.

Повернемося до кореляції, про яку ми говорили раніше. На діаграмі розсіювання (див. рис.4) сфокусуємося на одній точці та її відстані до лінії регресії та до її середнього значення.

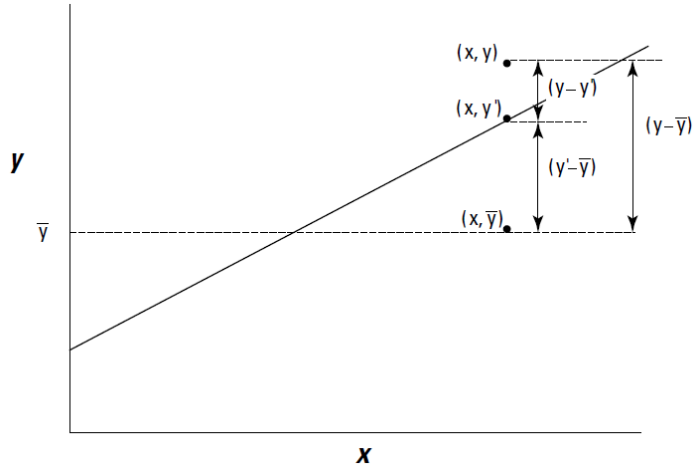


Рис.4. Одна точка на діаграмі розсіювання та пов'язані з нею відстані.

Зверніть увагу на три відстані, зображені на рисунку. Відстань, позначена (*y*-*y*'), є різницею між точкою та прогнозом лінії регресії, де ця точка має бути. Це називається нев'язкою (*residual*).) Відстань позначена $y\text{-}\left(\bar{y}\right)$ — це різниця між точкою та середнім значенням *y*.

Відстань позначена $y'\text{-}\left(\bar{y}\right)$ — це виграш у можливості прогнозування, який ми отримуємо від використання лінії регресії для прогнозування точки замість використання середнього значення.

На рис.4 показано, що відстань між точкою та лінією регресії та відстань між лінією регресії та середнім значенням дорівнюють відстані між точкою та середнім значенням:

$$\left(y-y'\right)+\left(y'-\bar{y}\right)=\left(y-\bar{y}\right)$$

Піднесемо кожне відхилення до квадрату. Це дасть нам наступне: $\left(y-y'\right)^2, \left(y'-\bar{y}\right)^2,$ і $\left(y-\bar{y}\right)^2.$

Якщо ми просумуємо кожне з квадратів відхилень, ми отримаємо:

$\sum(y-y')^2$. Це чисельник залишкової дисперсії. Він представляє мінливість навколо лінії регресії — «помилку». Чисельник дисперсії називається сумою квадратів, і часто позначається SS. Отже, це - $SS_{Residual}$.

$\sum(y'-\bar{y})^2$. Відхилення $\left(y'-\bar{y}\right)$ є виграшем у передбаченні завдяки використанню лінії регресії, а не середнього значення. Сума відображає цей приріст і називається $SS_{Regression}$.

$\sum(y-\bar{y})^2$. Це нумератор дисперсії *y*, або нумератор загальної дисперсії SS_{Total}.

Між цими трьома сумами існує залежність:

$$SS_{Residual} + SS_{Regression} = SS_{Total}$$

Кожна сума пов'язана зі значеннями ступенів свободи — знаменником оцінки дисперсії. Знаменник для $SS_{Residual}$ дорівнює N-2. df для SS_{Total} становить N-1. Як і у випадку з SS, ступені свободи сумуються:

$$df_{Regression} + df_{Residual} = df_{Total}$$

Якщо $SS_{Regression}$ є великим порівняно з $SS_{Residual}$, це означає, що зв'язок між змінною x і змінною y є сильним. Це означає, що по всій діаграмі розсіювання мінливість навколо лінії регресії є невеликою.

З іншого боку, якщо $SS_{Regression}$ є малою у порівнянні з $SS_{Residual}$, це означає, що зв'язок між x-змінною та y-змінною є слабкою. У цьому випадку мінливість навколо лінії регресії велика по всій діаграмі розсіювання.

Один із способів перевірити $SS_{Regression}$ проти $SS_{Residual}$ полягає в тому, щоб поділити кожну суму на її ступінь свободи (1 для $SS_{Regression}$ та N-2 для $SS_{Residual}$), щоб сформувати оцінки дисперсії (також відомі як середні квадрати, або MS), а потім розділити один на інший, щоб отримати F:

$$MS_{Regression} = \frac{SS_{Regression}}{df_{Regression}}$$

$$MS_{Residual} = \frac{SS_{Residual}}{df_{Residual}}$$

$$MS_{Total} = \frac{SS_{Total}}{df_{Total}}$$

$$F = \frac{MS_{Regression}}{MS_{Residual}}.$$

Якщо $MS_{Regression}$ значно перевищує $MS_{Residual}$, у нас є доказ того, що зв'язок x-y є сильним.

Ще один спосіб оцінити розмір $SS_{Regression}$ — порівняти його з SS_{Total}. Потрібно розділити перше на друге. Якщо співвідношення велике, це говорить про те, що зв'язок x-y є сильним.

Це співвідношення називається коефіцієнтом детермінації. Його символ — r^2. Якщо обчислюємо квадратний корінь із цього коефіцієнта, то ви отримаємо **коефіцієнт кореляції**.

$$r = \pm\sqrt{r^2} = \pm\sqrt{\frac{SS_{Regression}}{SS_{Total}}}$$

Знак «плюс» або «мінус» (±) означає, що r є додатним або від'ємним квадратним коренем залежно від того, позитивний чи від'ємний нахил лінії регресії.

Отже, якщо ми обчислюємо коефіцієнт кореляції і хочемо швидко дізнатися, що означає його значення, просто необхідно його піднести до квадрату. Коефіцієнт детермінації дає змогу дізнатися пропорцію SS_{Total}, пов'язаною із зв'язком між змінною x і змінною y. Якщо це велика пропорція, коефіцієнт кореляції означає сильний зв'язок. Якщо це невелика пропорція, коефіцієнт кореляції означає слабкий зв'язок.

У прикладі GPA-SAT коефіцієнт кореляції становить 0,817. Коефіцієнт детермінації становить:

$$r^2 = \left(.817\right)^2 = .667$$

У нашій вибірці з 20 студентів $SS_{Regression}$ становить 66,7 відсотка від SS_{Total}. Звучить як велика пропорція. Але що означає велика чи мала? Ці запитання вимагають перевірки гіпотез.

2.2. Перевірка гіпотези про кореляцію

Як і будь-який інший вид перевірки гіпотез, ідея полягає у використанні вибіркових статистичних даних, щоб зробити висновки щодо параметрів сукупності. Тут статистикою вибірки є *r* - коефіцієнт кореляції. За домовленістю параметр популяції дорівнює ρ (rho), грецький еквівалент *r*.

У зв'язку з кореляцією важливі два типи питань: (1) Чи коефіцієнт кореляції більший за нуль? (2) Чи відрізняються один від одного два коефіцієнти кореляції?

Повертаючись до прикладу SAT-GPA, ми можемо використовувати вибірку *r* для перевірки гіпотез щодо популяції ρ — коефіцієнта кореляції для всіх студентів університету.

Якщо припустити, що ми знаємо заздалегідь (до того, як ми зберемо будь-які вибіркові дані), що будь-яка кореляція між SAT і GPA має бути позитивною, гіпотези є наступними:

H_0: $\rho \leq 0$
H_1: $\rho > 0$

Встановимо $\alpha = .05$.

Відповідним статистичним тестом є t-тест. Його формула є наступною:

$$t = \frac{r - \rho}{s_r}$$

Цей тест має *N*-2 рівнів свободи df. Для нашого прикладу встановлено наступні значення у чисельнику: *r* дорівнює 0,817, а ρ (у H_0) дорівнює нулю. А як щодо знаменника? Це:

$$\sqrt{\frac{1-r^2}{N-2}}$$

З невеликою алгеброю формула для t-критерію спрощується до:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Наприклад,

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{.817\sqrt{20-2}}{\sqrt{1-.817^2}} = 6.011$$

При df = 18 і α = 0,05 (однобічний (one-tailed)), критичне значення t дорівнює 2,10 (В Excel для перевірки використовуйте функцію TINV). Оскільки розраховане значення перевищує критичне значення, приймаємо рішення відхилити H_0.

2.3. Приклад 1. Медицина

Встановимо, чи існує кореляційний зв’язок між масою тіла і артеріальним тиском? Оцініть характер та глибину (силу) кореляційного зв’язку, вірогідність коефіцієнту кореляції. Дані досліджень наведено в таблиці:

Маса, x_i	120	80	110	100	90
Артеріальний тиск, y_i	150	110	135	140	115

З вигляду кореляційного поля можна зробити припущення, що між ознаками X та Y існує пряма лінійна залежність, оскільки експериментальні точки групуються навколо прямої лінії. Наявність та тісноту лінійного зв'язку між двома ознаками визначимо за допомогою коефіцієнта кореляції Пірсона) *r*.

Обчислення за формулою наведеної вище зручно виконувати за допомогою наведеної нижче розрахункової таблиці, яка містить проміжні результати і відображає послідовність розрахунків.

X	Y	$\Delta=X_i-X$	$\Delta=Y_i-Y$	ΔX^2	ΔY^2	$\Delta X\cdot\Delta Y$
120	150	20	20	400	400	400
80	110	-20	-20	400	400	400
110	135	10	5	100	25	50
100	140	0	10	0	100	0
90	115	-10	-15	100	225	150
X=100	Y=130			$\Sigma\ \Delta X^2=1000$	$\Sigma\ \Delta Y^2=1150$	$\Sigma\ \Delta X\cdot\Delta Y=1000$

$$r_{x,y} = \frac{1000}{\sqrt{1000\cdot 1150}} = 0.9325,$$

Оскільки $r_{x,y} > 0$, то характер кореляційного зв'язку прямий.

Оскільки $0,9 < r_{x,y} < 1$, то глибина (сила) кореляційного зв'язку дуже сильна.

Оцінимо вірогідність коефіцієнта кореляції $r_{x,y}$:

а) $m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.9325^2}{5-2}} = 0,2085;$

б) $t_r = \frac{r}{m_r} = \frac{0.9325}{0.2085} = 4,4724;$

в) n = n – 2 = 5 – 2 = 3;

г) знаючи n, знайдемо за таблицею критерії Стьюдента t_{st} для трьох порогів вірогідності a_1, a_2, a_3: $t_{st} = \{t_{0,95}, \ldots t_{0,99}, \ldots t_{0.999}\}$ (див. додаток);

$t_{0,95}$ = 3,182,

$t_{0,99}$ = 5,840,

$t_{0.999}$ = 12,924.

г) оскільки $t_{0,95} < t_r < t_{0,99}$, то коефіцієнт кореляції достовірний.

2.4. Приклад 2. Дисперсія у фінансах.

Ось гіпотетичний приклад, щоб продемонструвати, як працює дисперсія. Припустимо, прибутковість акцій компанії ABC становить 10% у 1 рік, 20% у 2 рік і −15% у 3 рік. Середнє значення цих трьох доходів становить 5%. Різниця між кожною прибутковістю та середнім становить 5%, 15% і −20% для кожного наступного року.

Піднесення цих відхилень у квадрат дає 0,25%, 2,25% і 4,00% відповідно. Якщо ми додамо ці квадрати відхилень, то отримаємо 6,5%. Коли ми поділимо суму 6,5% на одиницю меншу за кількість об'єктів у наборі даних, оскільки це вибірка (2 = 3-1), це дає дисперсію 3,25% (0,0325). Взяття квадратного кореня з дисперсії дає стандартне відхилення 18% (√0,0325 = 0,180) для прибутків.

2.5. Функції для обчислення дисперсії та кореляції в Excel.

Є кілька різних варіантів формули для обчислення дисперсії в Excel:
=VAR.S(вибір даних)
=VARA(вибір даних)
=VAR.P(вибір даних)

Для кожної з цих формул потрібно вибрати діапазон комірок, який потрібно використовувати. Наприклад, можна ввести =VAR.S(B12:B32), щоб знайти дисперсію для даних у клітинках B12–B32.

=VAR.S(вибір даних)

Обчислює дисперсію на основі вибірки (ігноруючи логічні значення й текст у вибірці).

- Для функції VAR.S припускається, що її аргументи – це вибірка з генеральної сукупності. Якщо дані представляють генеральну сукупність, дисперсію слід обчислювати за допомогою функції VAR.P.
- Аргументи можуть бути числами, іменами, масивами або посиланнями, які містять числа.
- Логічні значення та числа у вигляді тексту, введені безпосередньо в списку аргументів, враховуються.
- Якщо аргумент — масив або посилання, враховуються лише числа в цьому масиві або посиланні. Пусті клітинки, логічні значення, текст, а також значення помилки в масиві або посиланні ігноруються.
- Аргументи, які являють собою значення помилок або текст, які не можна перетворити на числа, призводять до помилок.
- Якщо до посилання потрібно включити логічні значення та числа у вигляді тексту як частину обчислення, слід скористатися функцією VARA.
- У функції VAR.S використовується така формула:
$$\frac{\sum(x-\bar{x})^2}{(n-1)}$$
- де x – середнє значення вибірки AVERAGE(number1,number2,...) а n – розмір вибірки.

=VARA(вибір даних)

Оцінює дисперсію на основі вибірки.

- Для функції VARA припускається, що її аргументи є вибіркою із сукупності. Якщо дані представляють генеральну сукупність, дисперсію слід обчислювати за допомогою функції VARPA.
- Аргументами можуть бути: числами; іменами; масивами або посиланнями, які містять числа; числами у вигляді тексту; або логічними значеннями (TRUE та FALSE) у посиланні.
- Враховуються логічні значення та числа у вигляді тексту, введені безпосередньо у списку аргументів.
- Аргументи, які містять значення TRUE, інтерпретуються як 1; аргументи, які містять значення FALSE, інтерпретуються як 0 (нуль).
- Якщо аргумент є масивом або посиланням, використовуються лише значення з цього масиву або посилання. Пусті клітинки та текстові значення в масиві або посиланні ігноруються.
- Аргументи, що містять значення помилок або текст, який не можна перетворити на числа, спричиняють помилки.
- Якщо до посилання не слід включати логічні значення та числа у вигляді тексту як частину обчислення, скористайтеся функцією VAR.
- У функції VARA використовується така формула:
$$\frac{\sum(x-\bar{x})^2}{(n-1)}$$
- де x – середнє значення вибірки AVERAGE(значення1;значення2,...) а n – розмір вибірки.

=VAR.P(вибір даних)

Обчислює дисперсію на основі вибірки (ігноруючи логічні значення й текст у вибірці).

- Для функції VAR.S припускається, що її аргументи – це вибірка з генеральної сукупності. Якщо дані представляють генеральну сукупність, дисперсію слід обчислювати за допомогою функції VAR.P.
- Аргументи можуть бути числами, іменами, масивами або посиланнями, які містять числа.
- Логічні значення та числа у вигляді тексту, введені безпосередньо в списку аргументів, враховуються.
- Якщо аргумент — масив або посилання, враховуються лише числа в цьому масиві або посиланні. Пусті клітинки, логічні значення, текст, а також значення помилки в масиві або посиланні ігноруються.
- Аргументи, які являють собою значення помилок або текст, які не можна перетворити на числа, призводять до помилок.
- Якщо до посилання потрібно включити логічні значення та числа у вигляді тексту як частину обчислення, слід скористатися функцією VARA.
- У функції VAR.S використовується така формула:
$$\frac{\sum(x-\bar{x})^2}{(n-1)}$$
- де x – середнє значення вибірки AVERAGE(number1,number2,...) а n – розмір вибірки.

Excel надає дві функції для обчислення кореляції - CORREL і PEARSON. Це дві основні функції кореляції.

Функція CORREL повертає коефіцієнт кореляції двох діапазонів клітинок.

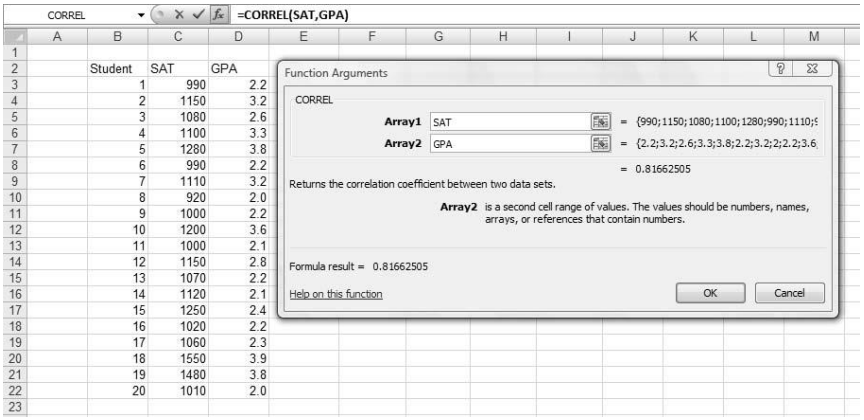
- Якщо аргумент масиву або посилання містить текст, логічні значення або пусті клітинки, ці значення ігноруються; Проте включаються клітинки з нульовими значеннями.
- Якщо масив1 і масив2 мають різну кількість точок даних, функція CORREL повертає помилку #N/A.
- Якщо масив1 або масив2 пустий або якщо **s** (стандартне відхилення) їхніх значень дорівнює нулю, функція CORREL повертає #DIV/0! помилку #REF!.

Функція PEARSON повертає коефіцієнт кореляції Пірсона r - безрозмірний індекс у діапазоні від -1,0 до 1,0 включно та відображає ступінь лінійного зв'язку між двома наборами даних.

- Аргументи мають бути або числами, або іменами, константами масиву або посиланнями, які містять числа.
- Якщо аргумент масиву або посилання містить текст, логічні значення або порожні клітинки, ці значення ігноруються; однак клітинки з нульовим значенням включені.
- Якщо масив1 і масив2 порожні або мають різну кількість точок даних, PEARSON повертає значення помилки #N/A.

Інші - RSQ та COVAR. RSQ обчислює коефіцієнт визначення (квадрат коефіцієнта кореляції), а COVAR обчислює коваріацію.

На рисунку показані дані для прикладу SAT-GPA, а також діалогове вікно «Аргументи функцій» для CORREL.



	Student	SAT	GPA
1	1	990	2.2
2	2	1150	3.2
3	3	1080	2.6
4	4	1100	3.3
5	5	1280	3.8
6	6	990	2.2
7	7	1110	3.2
8	8	920	2.0
9	9	1000	2.2
10	10	1200	3.6
11	11	1000	2.1
12	12	1150	2.8
13	13	1070	2.2
14	14	1120	2.1
15	15	1250	2.4
16	16	1020	2.2
17	17	1060	2.3
18	18	1550	3.9
19	19	1480	3.8
20	20	1010	2.0

Вибір PEARSON замість CORREL дає вам точно таку ж відповідь, і ви використовуєте її точно так само.

RSQ

Якщо потрібно швидко обчислити коефіцієнт визначення (r^2) , тоді потрібно використати функцію RSQ.

Ось як виглядає панель формул EXCEL після заповнення діалогового вікна «Аргументи функцій RSQ» для цього прикладу: **= RSQ (GPA, SAT)**

Що стосується діалогового вікна, то єдина відмінність між цим і CORREL (і PEARSON) полягає в тому, що поля, які ви заповнюєте, називаються Known_y's та Known_x's, а не Array 1 та Array 2.

COVAR

Ця функція використовується так само, як і CORREL. Після заповнення діалогового вікна «Аргументи функцій» у цьому прикладі формула на панелі формул є

= COVAR (SAT, GPA)

Якщо ви хочете використовувати цю функцію для обчислення r, потрібно поділити результат на добуток STDEVP (SAT) та STDEVP (GPA).

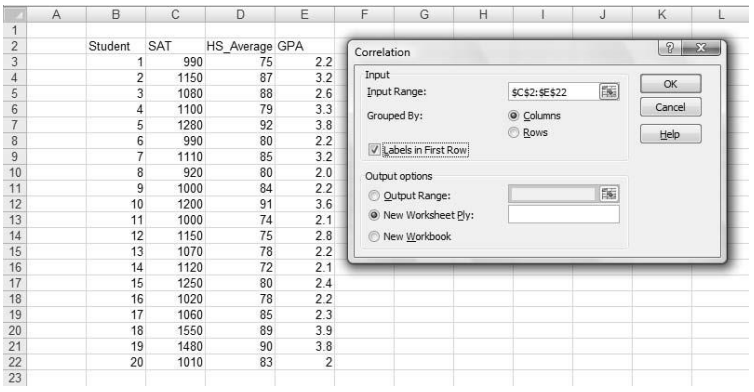
Замість цього можна просто використати функцію CORREL.

Якщо потрібно обчислити лише коефіцієнт кореляції, то можна застосувати інструмент аналізу даних кореляції Excel який робить те саме, що і CORREL, і виводить результат у табличній формі.

Цей інструмент стає корисним, коли потрібно обчислити кілька кореляцій на наборі даних.

Наприклад, на рисунку показано SAT, High School Average та GPA для 20 студентів університету, а також діалогове вікно інструменту аналізу даних про кореляцію.

На рисунку показано табличний результат інструменту аналізу даних кореляції - кореляційна матриця.



	A	B	C	D	E
1		*SAT*	*HS_Average*	*GPA*	
2	SAT	1			
3	HS_Average	0.552527329	1		
4	GPA	0.81662505	0.714353653	1	
5					

Множинна кореляція

	A	B	C	D	E
1		*SAT*	*HS_Average*	*GPA*	
2	SAT	1			
3	HS_Average	0.552527329	1		
4	GPA	0.81662505	0.714353653	1	
5					

Для цього прикладу,

$$R_{G.SH} = \sqrt{\frac{(.816625)^2 + (.714354)^2 - 2(.816625)(.714354)(.552527)}{1-(.816625)^2}} = .875529$$

Якщо піднести до квадрату це число, ми отримаємо множинний коефіцієнт кореляції.

$$R^2_{G.SH} = (.875529)^2 = .766552$$

Часткова кореляція

GPA та SAT асоціюються із High School Average (у прикладі). Кожна асоціація із High School Average може якось приховати справжню кореляцію між ними.

Яким буде їх співвідношення, якби ми могли усунути цю асоціацію?

Іншими словами, якою була б кореляція GPA-SAT, якби ми могли тримати значення High School Average постійним?

Один із способів утримувати постійну High School Average- це знайти кореляцію GPA-SAT для вибірки студентів, у яких High School Average, наприклад, - 87. У подібній вибірці співвідношення кожної змінної із High.

School Average дорівнює нулю. Проте в реальному світі це неможливо.

Інший спосіб - знайти часткову кореляцію між GPA та SAT. Це статистичний спосіб усунення асоціації кожної змінної із High School Average у нашій вибірці. Для цього використаємо коефіцієнти кореляції в кореляційній матриці: $r_{GS.H} = \frac{r_{GS} - r_{GH}r_{SH}}{\sqrt{1-r^2_{GH}}\sqrt{1-r^2_{SH}}}$

G означає GPA, S - SAT, а H - High School Average. Підписка GS.H означає, що кореляція між GPA та SAT із High School Average "частково відключена".

Для цього прикладу: $r_{GS.H} = \frac{.816625 - (.714353)(.552527)}{\sqrt{1-(.714353)^2}\sqrt{1-(.552527)^2}} = .547005$

Засіб аналізу даних: Covariance

Можна використати інструмент аналізу даних Коваріації так само, як і Інструмент аналізу кореляційних даних. Табличний результат представлено на рисунку.

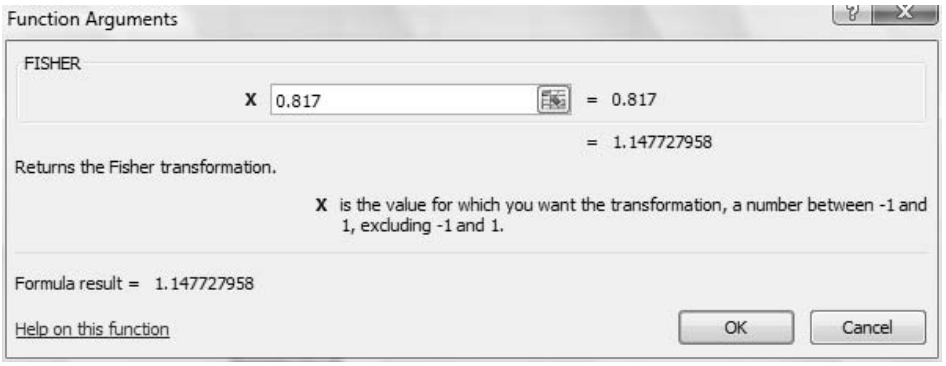
Таблиця - коваріаційна матриця. Кожна комірка в матриці показує коваріацію змінної в рядку зі змінною в стовпці (знову ж таки, використовуючи N, а не N-1). Клітинка C4 показує коваріацію GPA із High School Average. Основна діагональ у цій матриці представляє дисперсію кожної змінної (що еквівалентно коваріації змінної із самою собою). У цьому випадку дисперсія - це те, що обчислюється із функцією VARP.

	A	B	C	D	E
1		*SAT*	*HS_Average*	*GPA*	
2	SAT	24862.75			
3	HS_Average	512.375	34.5875		
4	GPA	85.1675	2.77875	0.437475	
5					

Функції: FISHER, FISHERINV

Excel обробляє досить складні перетворення, які дозволяють перевірити гіпотези про різницю двох коефіцієнтів кореляції. FISHER перетворює r в z. FISHERINV робить зворотне.

Ми використовуємо перетворені значення у формулі, $Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}}$ в якій знаменник $\sigma_{z_1-z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$



4. ЛАБОРАТОРНЕ ЗАВДАННЯ

Підготуйте дані для завантаження у Excel. Використовуючи можливості табличного процесору Excel:

1. обчислити коефіцієнт кореляції, перевірити його значимість і надійність;
2. обчислити параметри прямого і оберненого прогнозів;
3. побудувати графіки прогнозів;
4. визначити кут між графіками прямого та оберненого прогнозів;
5. на основі отриманих результатів зробити висновки.

Індивідуальні завдання:

Варіант 1

№	y	x_1	x_2
1	82	18,9	43
2	70	22,8	54
3	66	23,1	48,8
4	60	22,8	57,2
5	50	27,3	44,2
6	38	32,4	65
7	37	31,5	65
8	24	37	73,6
9	20	39	77
10	19	39	83
11	17,5	42,7	62,2
12	16	42	64

Варіант 2

№	y	x_1	x_2
1	182	18,9	40
2	162	24,8	43
3	140	29	60
4	125	37	64
5	110	46,8	69,2
6	98	44,4	74,8
7	78	37,8	91
8	66	49	80
9	46	48,8	93,6
10	34	58	93,4
11	14	58,8	107
12	12	58	110

Варіант 3

№	y	x_1	x_2
1	24	25	12,5
2	26,8	31	19,2
3	28,2	34	21,2
4	29,6	57	37,4
5	31	40	48
6	32,4	43	55,6
7	33,8	46	56,8
8	35,2	69	73,8
9	36,6	52	83,6
10	35,9	55	92
11	37	60	80
12	38	64	85

Варіант 4

№	y	x_1	x_2
1	11	30	12,5
2	16	32	20
3	19	36	19,5
4	30	41	40
5	33	42	44,5
6	40	55	57
7	47	48	58
8	65	52	80
9	66	54	79
10	70	55	80
11	75	42,7	81
12	80	42	85

Варіант 5

№	y	x_1	x_2
1	24	25	12,5
2	26,8	31	19,2
3	34	9	25,5
4	33,2	7,6	17,1
5	40,4	10,8	29,8
6	39,6	9,2	37,3
7	46,8	12,6	54,6
8	47	10,4	55,5
9	53,2	14,4	57,4
10	52,4	12	55,7
11	59,6	16,2	68,2
12	58,8	13,2	79,9

Варіант 6

№	y	x_1	x_2
1	21	30	12,5
2	24	35	25,5
3	26,8	51	17,2
4	28,2	44	29,7
5	29,6	54,5	37,4
6	31	50	54,5
7	32,4	45	55,6
8	33,8	56	57,3
9	35,2	61	73,8
10	36,6	62	68,1
11	35,9	60	80
12	38	64	85

Варіант 7

№	y	x_1	x_2
1	8	3,6	25,5
2	8,8	36,2	17,2
3	9,2	4,5	29,7
4	9,6	1,3	37,4
5	10	5,1	54,5
6	10,4	45	55,6
7	10,8	5,7	57,3
8	9,9	6	80
9	10	6,3	68,1
10	11,2	61	73,8
11	11,2	61	73,8
12	11,6	6,3	68,1

Варіант 8

№	y	x_1	x_2
1	34	21,4	25,5
2	36,4	20,6	17,2
3	46,8	29,8	29,6
4	49,2	35	37,6
5	59,6	38,2	54,2
6	63	32,6	56
7	72,4	46,6	56,8
8	74,8	50,2	56,4
9	85,2	55	67,4
10	87,6	47,6	80,8
11	90	61	73,8
12	92	63	68,1

Варіант 9

№	y	x_1	x_2
1	145	21,4	25,5
2	163	20,6	17,2
3	202	31,4	19
4	212	32,6	9,2
5	282	43,8	14,1
6	330	51	21,6
7	344	56,2	37,7
8	399	52,6	32
9	442	68,6	25,3
10	478	74,2	24,4
11	524	81	34,9
12	525	75,6	40,8

Варіант 10

№	y	x_1	x_2
1	202	31,2	106
2	218	29,3	102,2
3	242	37,4	98,6
4	258	42,5	105,2
5	282	43,6	112
6	299	38,3	109
7	322	49,8	106,2
8	338	54,1	113,6
9	362	56,1	121,2
10	378	49,5	119
11	380	100,7	90
12	400	102	95

Варіант 11

№	y	x_1	x_2
1	95	44	80
2	101	47	100
3	109	58,3	97,8
4	121	56,4	93,4
5	129	68,9	94,8
6	141	65,8	93
7	150	79,5	91
8	161	75,2	84,8
9	169	90,1	86,4
10	181	94,6	85,8
11	189	100,7	90
12	190	102	91

Варіант 12

№	y	x_1	x_2
1	150	15	106
2	182	17	89,8
3	221	25,3	87,8
4	218	20,4	83,6
5	250	29,9	84,8
6	254	23,8	82,8
7	270	34,5	81
8	290	27,2	75
9	305	39,1	76,4
10	326	40,6	75,6
11	341	43,7	80
12	400	102	95

Варіант 13

№	y	x_1	x_2
1	126	21	93
2	182	54	89,8
3	162	51,2	104
4	174	46,3	97,8
5	194	61,4	91,4
6	206	73,5	94,8
7	226	71,6	98
8	239	63,3	91
9	258	81,8	83,8
10	270	93,1	86,4
11	290	92,1	88,8
12	302	82,5	85

Варіант 14

№	y	x_1	x_2
1	24	17	69,4
2	26	25,3	67,8
3	29	20,4	64
4	31	29,9	64,8
5	34	23,8	62,4
6	36	34,5	61
7	39	27,2	55,4
8	41	39,1	56,4
9	43	40,6	55,2
10	45	43,7	60
11	51	69	88,8
12	58	82,5	85

Варіант 15

№	y	x_1	x_2
1	49	17	31
2	48,5	16	37
3	48	16	46
4	47,6	15	52
5	47,3	12	71
6	46,9	11	91
7	46,5	11,5	92
8	46,3	11	145
9	46,1	10	190
10	45,9	10	204
11	45,7	9,5	222
12	45,6	8	271

Варіант 16

№	y	x_1	x_2
1	50	43	15
2	44	41	14,8
3	47	43,8	10,6
4	41	36,4	6,6
5	45	43,4	14,8
6	36	31,8	23,2
7	38	38	19,8
8	31	27,2	16,6
9	36	36,6	25,6
10	28	28	34,8
11	26	30,2	37
12	28	30	35

Варіант 17

№	y	x_1	x_2
1	249	17	91,9
2	237	15,3	92,2
3	219	20,4	92,7
4	207	39,9	95,2
5	189	23,8	97,9
6	177	24,5	99
7	159	27,2	100,3
8	147	49,1	103,6
9	129	30,6	107,1
10	117	33,7	109
11	108	41	109
12	100	30	112

Варіант 18

№	y	x_1	x_2
1	47	44	5,6
2	50	43,3	5
3	41	36,4	6,6
4	44	43,9	8,4
5	35	31,8	10,4
6	38	36,5	12,6
7	29	27,2	13,8
8	32	37,1	19,6
9	23	22,6	22,4
10	26	29,7	23,4
11	27	33	26
12	30	30	32

Варіант 19

№	y	x_1	x_2
1	27	44	8,2
2	36	43,8	7,8
3	33	36,4	7,6
4	42	43,4	11,6
5	39	31,8	15,8
6	48	37	16,2
7	45	27,2	16,8
8	54	36,6	21,6
9	51	22,6	26,6
10	60	30,2	27,8
11	62	33	29
12	65	34	32

Варіант 20

№	y	x_1	x_2
1	27	16,8	164,1
2	35	15,5	197
3	30	15,2	187,5
4	40	14,9	218,6
5	35	11,6	230,1
6	46	12,3	240,6
7	39	12	241
8	52	11,7	272,8
9	48	10,4	278
10	57	10	305
11	55	9	300
12	51	9	304

Варіант 21

№	y	x_1	x_2
1	270	16,8	152
2	350	15,5	197
3	414	2,1	520
4	468	1,9	450
5	498	1,8	610
6	552	1,9	720
7	582	1,4	750
8	636	1,5	770
9	666	1,4	850
10	720	1,6	830
11	750	1,3	990
12	804	1,1	980

Варіант 22

№	y	x_1	x_2
1	78	43	57,4
2	90	61,1	63,2
3	94	59,8	73,8
4	106	70,7	72,8
5	110	51	76
6	122	87,7	80,4
7	126	52,6	90,6
8	138	107,5	88
9	142	72,6	92,6
10	154	85	95,6
11	159	89	99
12	204	145	98

Варіант 23

№	y	x_1	x_2
1	7,2	31	15
2	7,9	45	16
3	9,2	61	19
4	9,4	72	18,2
5	11	75	17
6	11,3	77	18
7	11,5	85	13,9
8	12,3	83	15,4
9	13	99	11,7
10	13,8	98	13,2
11	14	95	15
12	17	98	22

Варіант 24

№	y	x_1	x_2
1	50	17	93
2	47	15,3	105
3	41	20,4	91,7
4	34	29,9	91,6
5	33	23,8	91,4
6	32	24,5	90
7	30	27,2	86
8	26	39,1	91
9	25	30,6	83,8
10	24	33,7	86,4
11	21	41	86,2
12	20	30	74

Варіант 25

№	y	x_1	x_2
1	59	43	28,2
2	67	61	29,6
3	71	59,8	35,4
4	79	70,8	34,4
5	83	51	38
6	91	87,6	37,2
7	95	52,6	41,8
8	103	107,6	40
9	107	72,6	45,8
10	115	112,6	42,8
11	141	95	52
12	172	98	54,1

Варіант 26

№	y	x_1	x_2
1	59	61	12
2	67	57,8	9,2
3	77	66	7,8
4	86	77,2	8,8
5	89	78	10
6	91	75	8,4
7	100	85	6,6
8	108	92,5	8
9	114	102	10
10	115	97	7
11	116	95	6
12	117	98	5

Варіант 27

№	y	x_1	x_2
1	59	12,6	9,4
2	67	12,9	9,4
3	71	15,4	9,6
4	79	18,1	10
5	83	18,2	10,6
6	91	18,1	11,4
7	95	21	12,4
8	103	24,1	13,6
9	107	23,8	15
10	115	23,3	16,6
11	116	25	18
12	59	12,6	9,4

Варіант 28

№	y	x_1	x_2
1	57	18,9	42
2	50	16,6	50,2
3	42	18,1	59,6
4	39	21,4	57,2
5	37	27,3	54
6	30	29,6	65
7	19	31,5	77,2
8	17	38,4	73,6
9	16	35,7	69,2
10	15	30,2	83
11	9	25	99
12	8	33,6	88

Варіант 29

№	y	x_1	x_2
1	62	18,9	43
2	56	22,8	50,2
3	56	23,1	48,8
4	54	22,8	57,2
5	54	27,3	44,2
6	42	32,4	65
7	34	41	65
8	29,8	46,8	73,6
9	28	48,7	62,2
10	21	52	83
11	20	51	89
12	15	64	102

Варіант 30

№	y	x_1	x_2
1	101	18,9	43
2	89	18,4	50,2
3	72	23,1	58,4
4	69	28	57,2
5	68	27,3	55,4
6	59	26,4	65
7	44	31,5	75,6
8	29	43,8	73,6
9	28	42,7	71
10	19	44,4	83
11	9	48	86,2
12	8	52	78

5. КОНТРОЛЬНІ ЗАПИТАННЯ

1. Які види зв'язків між явищами ви знаєте? Дайте визначення і коротку характеристику.
2. В чому сутність кореляційного зв'язку?
3. Які бувають зв'язку за напрямком? Як вони виражаються?
4. Які методи застосовуються в статистики для встановлення зв'язку між явищами, в чому їх суть? Навести приклад.
5. Назвіть основні задачі кореляційного аналізу і варіанти кореляційного зв'язку.
6. Що розуміють під рівнянням зв'язку і як визначаються його параметри?
7. Що таке щільність зв'язку і як вона визначається для різних форм зв'язку?
8. Як засобами Excel розрахувати коефіцієнт кореляції?
9. Які властивості має коефіцієнт кореляції?

10. Як встановити значимість коефіцієнта кореляції?
11. Як перевірити надійність коефіцієнта кореляції?

6. ЗМІСТ ЗВІТУ

1. Тема і мета роботи.
2. Завдання до роботи.
3. Результати виконання завдань.
4. Відповіді на контрольні запитання.
5. Висновки, що відображують результати виконання роботи та їх критичний аналіз.

Таблиця значень критерія Стьюдента

(t-критерій)

Критичні значення коєфіцієнта Стьюдента (t-критерій) для різних довірчих ймовірностей *p* та числа ступенів свободи *f*:

f	*p*							
	0.80	0.90	0.95	0.98	0.99	0.995	0.998	0.999
1	3.0770	6.3130	12.7060	31.820	63.656	127.656	318.306	636.619
2	1.8850	2.9200	4.3020	6.964	9.924	14.089	22.327	31.599
3	1.6377	2.35340	3.182	4.540	5.840	7.458	10.214	12.924
4	1.5332	2.13180	2.776	3.746	4.604	5.597	7.173	8.610
5	1.4759	2.01500	2.570	3.649	4.0321	4.773	5.893	6.863
6	1.4390	1.943	2.4460	3.1420	3.7070	4.316	5.2070	5.958
7	1.4149	1.8946	2.3646	2.998	3.4995	4.2293	4.785	5.4079
8	1.3968	1.8596	2.3060	2.8965	3.3554	3.832	4.5008	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.780
10	1.3720	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.363	1.795	2.201	2.718	3.105	3.496	4.024	4.437
12	1.3562	1.7823	2.1788	2.6810	3.0845	3.4284	3.929	4.178
13	1.3502	1.7709	2.1604	2.6503	3.1123	3.3725	3.852	4.220
14	1.3450	1.7613	2.1448	2.6245	2.976	3.3257	3.787	4.140
15	1.3406	1.7530	2.1314	2.6025	2.9467	3.2860	3.732	4.072
16	1.3360	1.7450	2.1190	2.5830	2.9200	3.2520	3.6860	4.0150
17	1.3334	1.7396	2.1098	2.5668	2.8982	3.2224	3.6458	3.965
18	1.3304	1.7341	2.1009	2.5514	2.8784	3.1966	3.6105	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.3253	1.7247	2.08600	2.5280	2.8453	3.1534	3.5518	3.8495
21	1.3230	1.7200	2.2.0790	2.5170	2.8310	3.1350	3.5270	3.8190
22	1.3212	1.7117	2.0739	2.5083	2.8188	3.1188	3.5050	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.1040	3.4850	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.0905	3.4668	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.4502	3.7251
26	1.315	1.705	2.059	2.478	2.778	3.0660	3.4360	3.7060
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.4210	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.0469	3.4082	3.6739
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.0360	3.3962	3.8494
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.0298	3.3852	3.6460
32	1.3080	1.6930	2.0360	2.4480	2.7380	3.0140	3.3650	3.6210
34	1.3070	1.6909	2.0322	2.4411	2.7284	3.9520	3.3479	3.6007