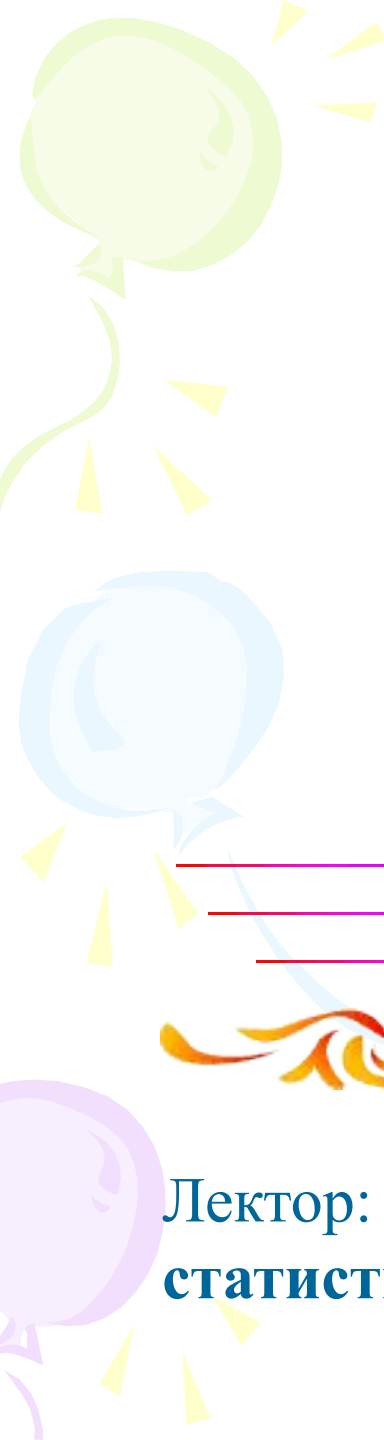


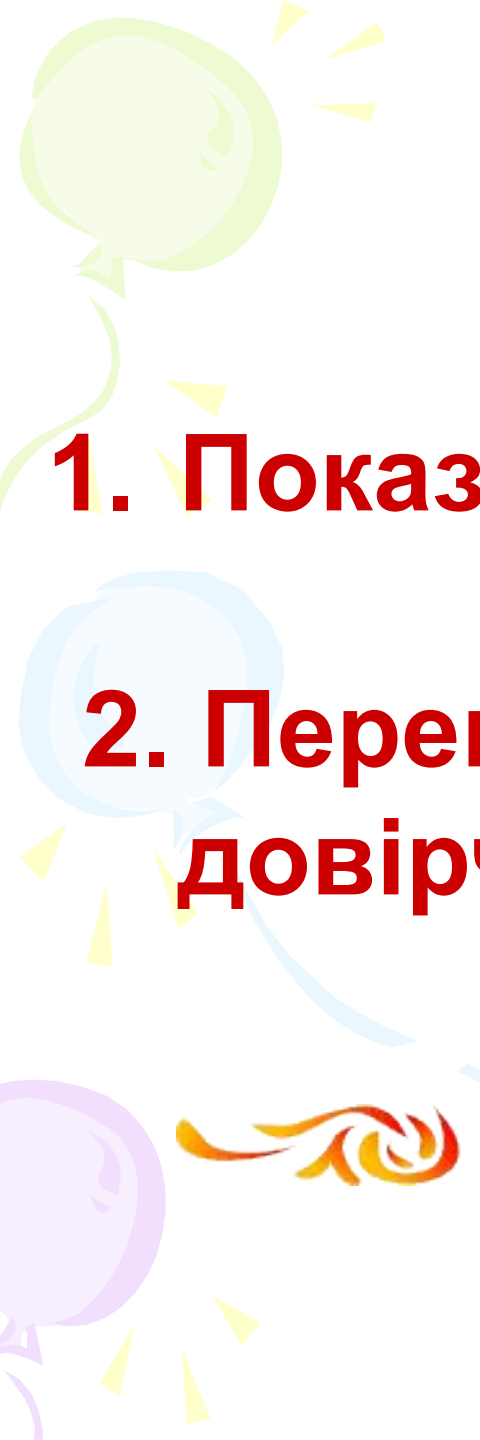


Тема 3

Верифікація моделі



Лектор: к.е.н., доцент кафедри економетрії та статистики ДЕМЧИШИН М.Я.



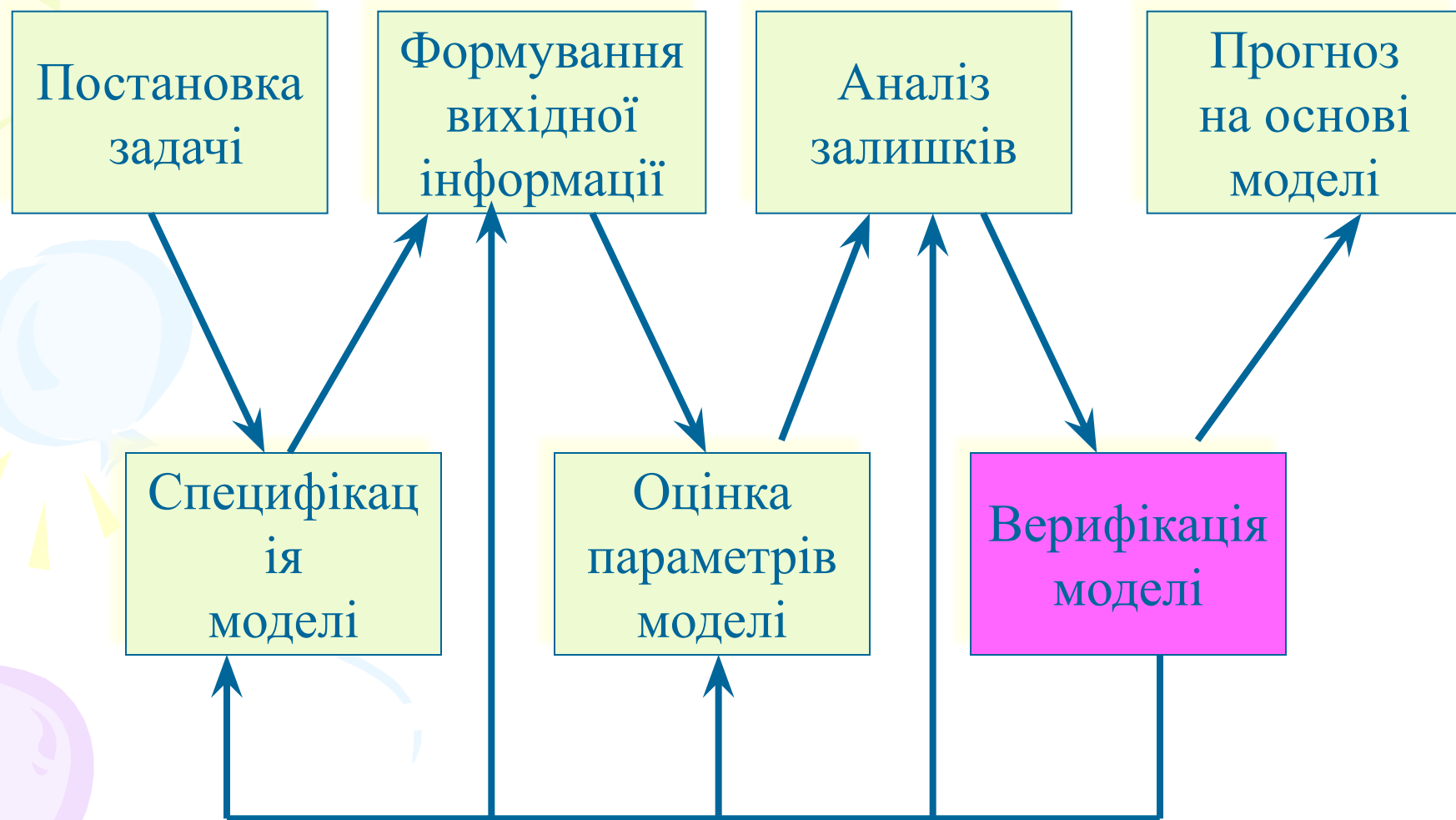
Зміст

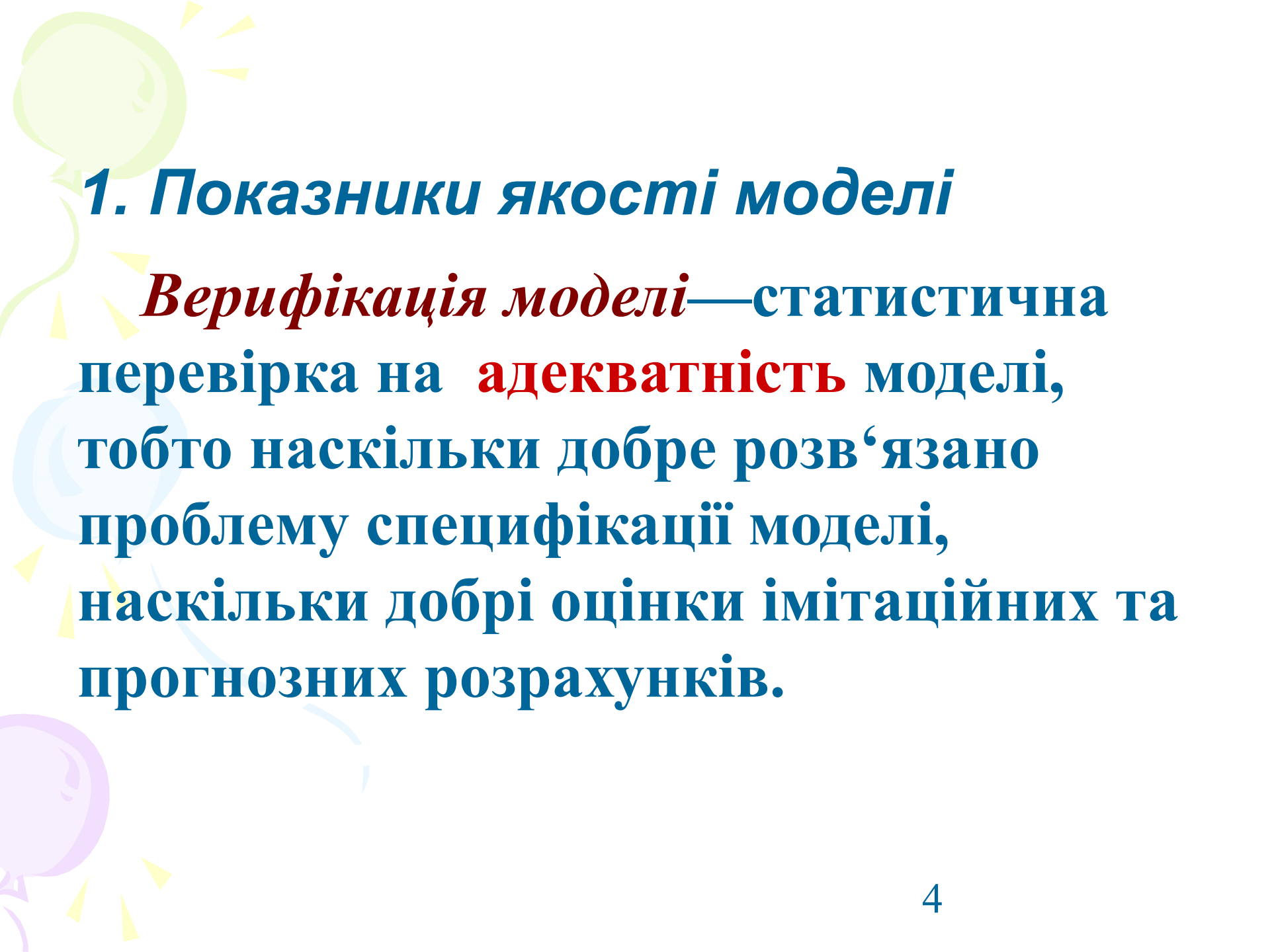
1. Показники якості моделі

**2. Перевірка значущості та
довірчі інтервали**



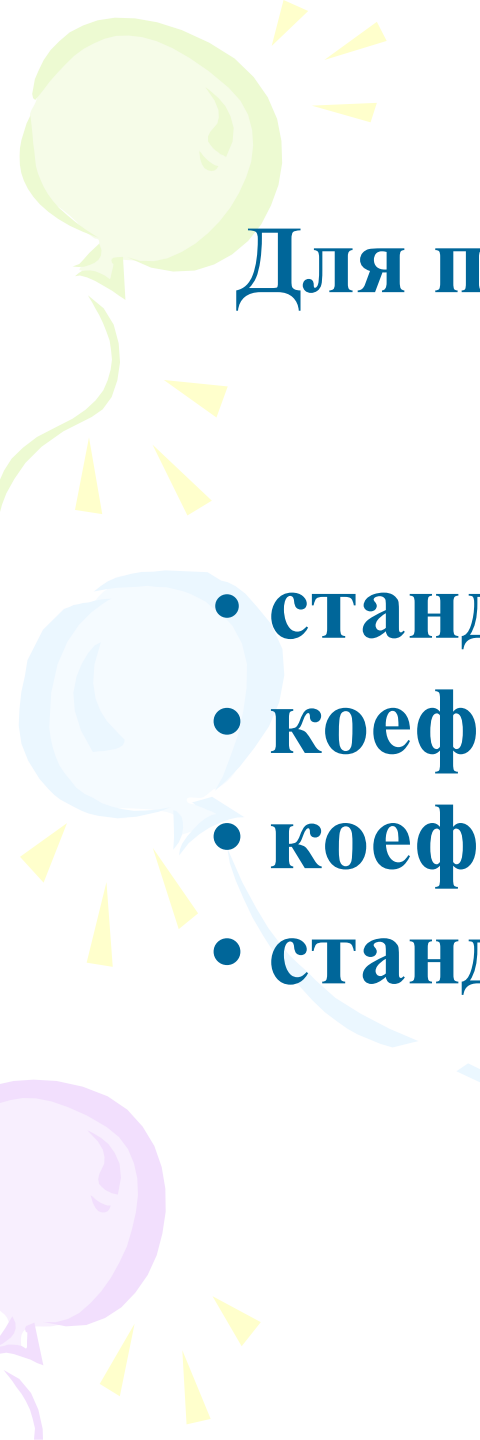
Етапи побудови моделі





1. Показники якості моделі

Верифікація моделі—статистична перевірка на **адекватність** моделі, тобто наскільки добре розв'язано проблему специфікації моделі, наскільки добрі оцінки імітаційних та прогностичних розрахунків.

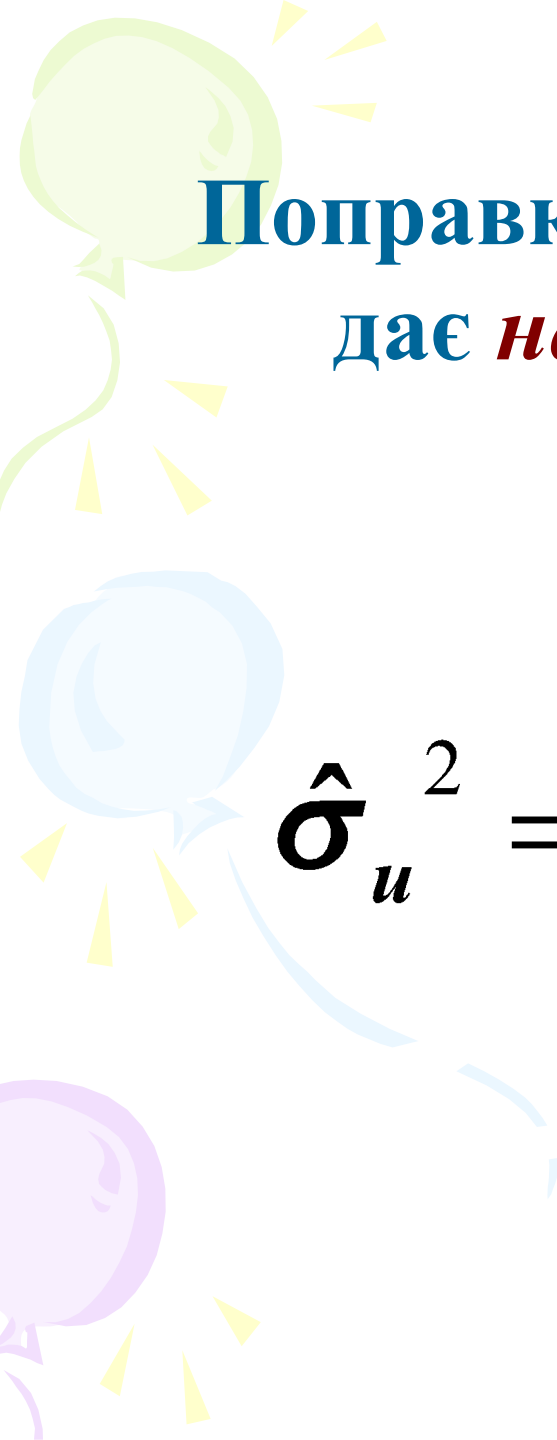


**Для перевірки коректності побудови
моделі визначають:**

- **стандартну похибку рівняння;**
- **коефіцієнт детермінації;**
- **коефіцієнт множинної кореляції;**
- **стандартну похибку параметрів.**

Стандартна похибка рівняння (точкова оцінка емпіричної дисперсії залишків)- характеризує абсолютну величину розкиду випадкової складової рівняння

$$S_u^2 = \frac{1}{n} \sum_{i=1}^n u_i^2$$



Поправка на число ступенів свободи
дає *незміщену оцінку* дисперсії
залишків:

$$\hat{\sigma}_u^2 = \frac{1}{n - m - 1} \sum_{i=1}^n u_i^2$$



У поняття "тіснота зв'язку" (*щільність*) вкладається оцінка впливу незалежної змінної на залежну змінну.



Під терміном "значимість зв'язку" (*істотність, або значущість*) розуміють оцінку відхилення вибіркових змінних від своїх значень у генеральній сукупності спостережень за допомогою статистичних критеріїв.

Коефіцієнт детермінації показує, якою мірою варіація залежної змінної (результативного показника) ***y*** визначається варіацією незалежної змінної (вхідного показника) ***x***.

$$R^2 = 1 - \frac{S_u^2}{S_y^2}$$

де

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Інші формули:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y} - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

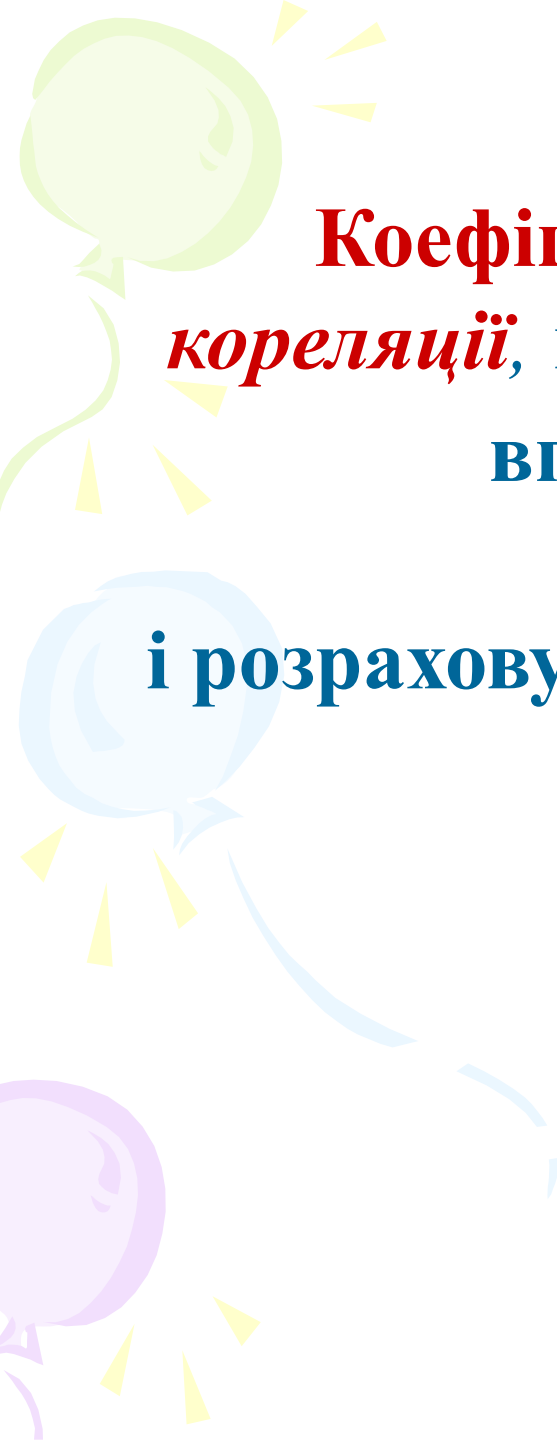
де $\hat{y}$ - розрахункові значення регресанда;

$\bar{y}$ -загальна середня фактичних
даних результативного показника;

y_i -фактичні індивідуальні значення
результативного показника.

$$R^2 = \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}) \right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$$

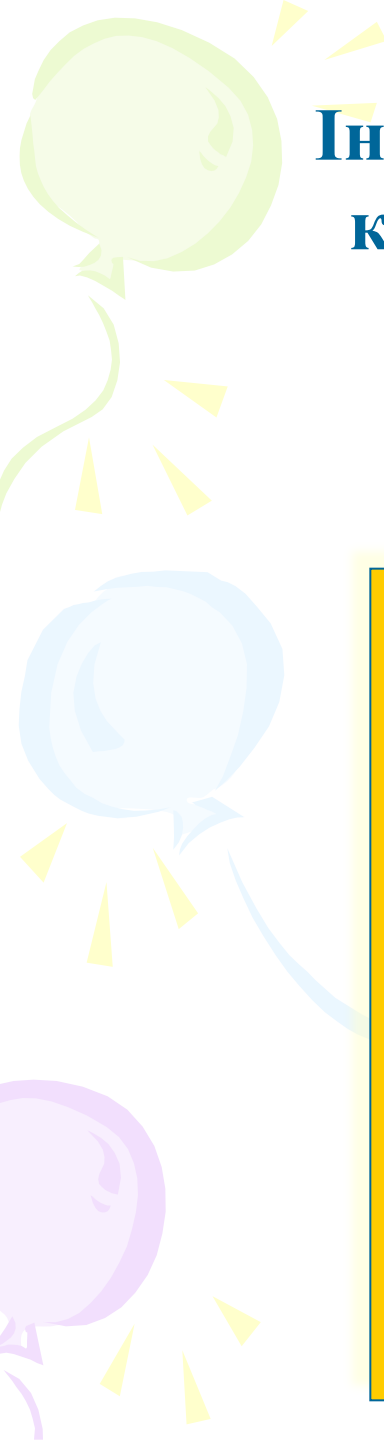
квадрат емпіричного коефіцієнта кореляції між
двома рядами спостережень (теоретичними
значеннями регресанта y_i та його
розрахунковими значеннями $\hat{y}$).



Коефіцієнт кореляції, або *індекс кореляції*, показує, наскільки значним є вплив змінної x_i , на y_i

і розраховується так:

$$R = \sqrt{R^2}$$



**Іноді для спрощення розрахунків тісноту
кореляційного зв'язку характеризують
коефіцієнтом кореляції, який
розраховується за формулою:**

$$R_1 = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Якщо зв'язок між результативним і вхідним показниками лінійний, то використовується *лінійний коефіцієнт кореляції*, який характеризує не тільки тісноту зв'язку, а і його напрям:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}$$

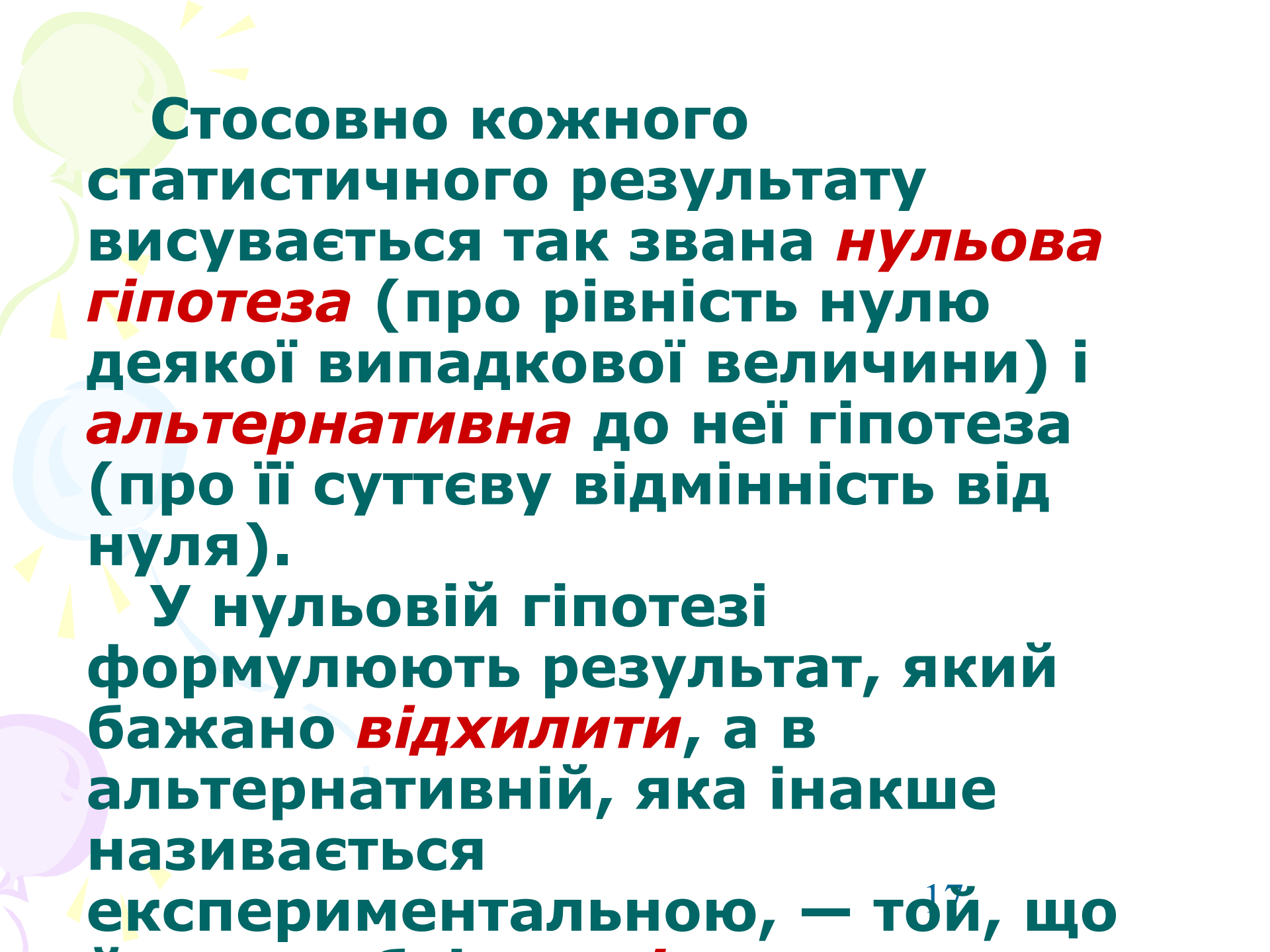


2. Перевірка значущості та довірчі інтервали.



Зауваження. У задачах регресійного аналізу важливе значення має припущення про **нормальний розподіл** випадкових величин, що задіяні в даній моделі.

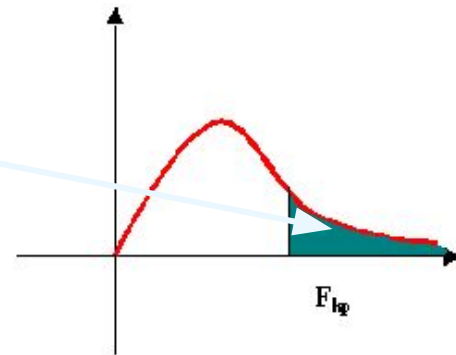
Певні перетворення нормально розподілених величин забезпечують їх розподіл за законом **Стюдента** чи за законом **Фішера**: на підставі першого з них визначаються **довірчі інтервали**, а другий дає змогу оцінювати відношення двох випадкових величин.

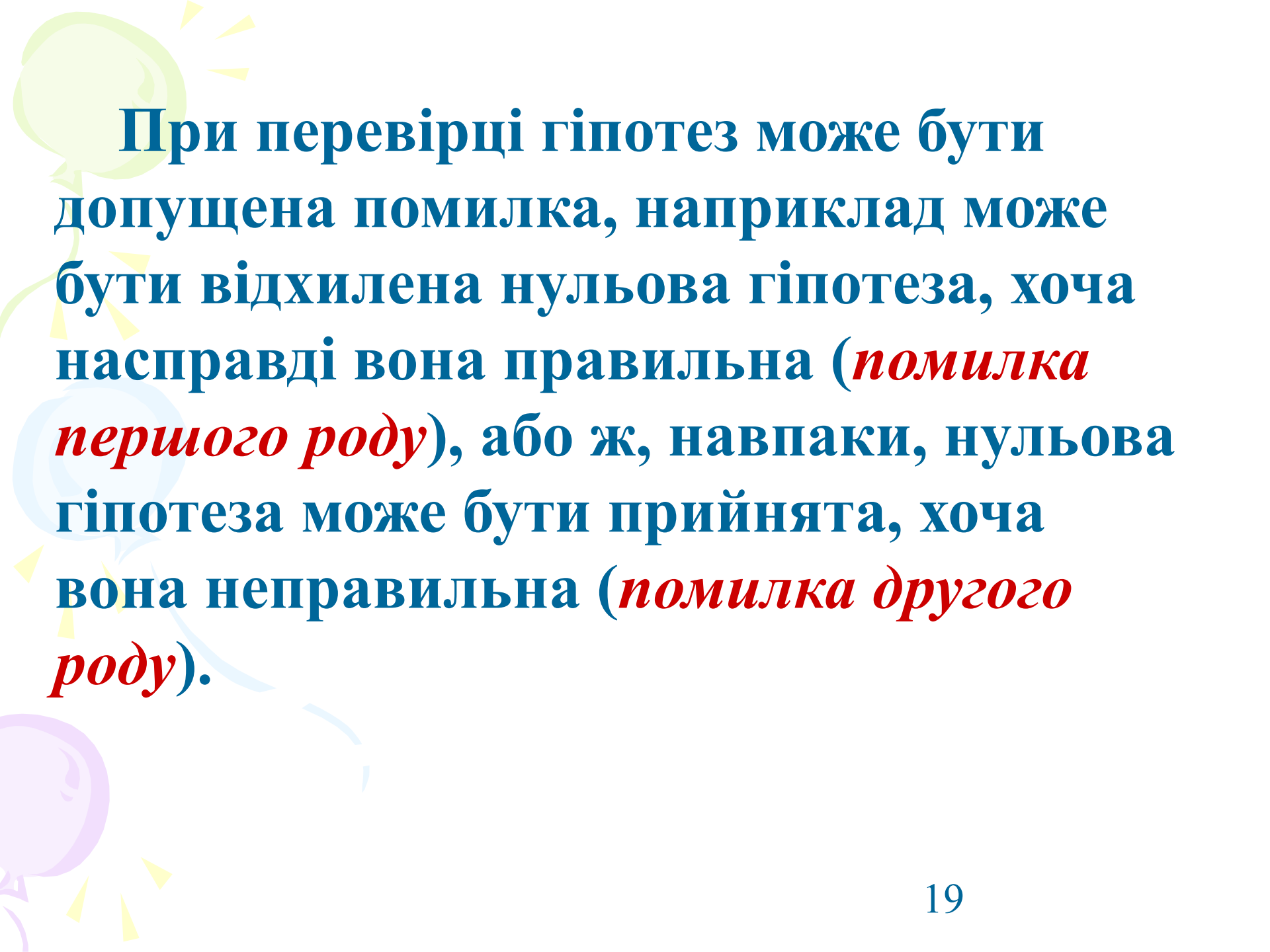


Стосовно кожного статистичного результату висувається так звана **нульова гіпотеза** (про рівність нулю деякої випадкової величини) і **альтернативна** до неї гіпотеза (про її суттєву відмінність від нуля).

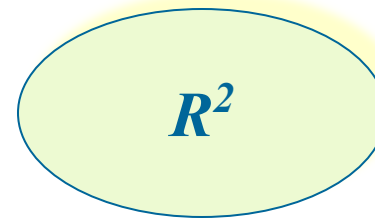
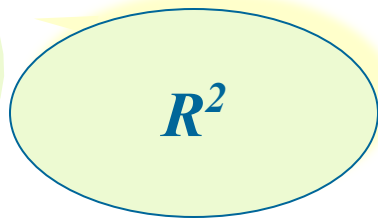
У нульовій гіпотезі формулюють результат, який бажано **відхилити**, а в альтернативній, яка інакше називається експериментальною, — той, що

За заданим *рівнем значущості* множина допустимих значень розбивається на дві неперетинні множини: одна містить значення випадкової величини, ймовірність досягнення яких перевищує заданий рівень значущості, а інша — *критична область* — визначає ті значення, що досягаються рідко (ймовірність потрапити до такої області нижча від заданого рівня), і розташована вона, як правило, на "хвостах розподілу".





При перевірці гіпотез може бути допущена помилка, наприклад може бути відхилена нульова гіпотеза, хоча насправді вона правильна (*помилка першого роду*), або ж, навпаки, нульова гіпотеза може бути прийнята, хоча вона неправильна (*помилка другого роду*).



?

**застосувати відповідний статистичний критерій,
який дасть змогу встановити,
чи суттєво відрізняється R^2 від нуля,
чи ця відмінність пов'язана з особливостями
конкретних даних,
тобто зумовлена лише похибками вимірювань.**

Висувається нульова гіпотеза $H_0 : R^2 = 0$.

Це означає, що досліджуване рівняння **не пояснює змінювання регресанда** під впливом відповідних регресорів.

У такому разі всі коефіцієнти при незалежних змінних мають дорівнювати нулю.

При цьому нульову гіпотезу можна подати у вигляді

$$H_0 : a_1 = a_2 = \dots = a_n = 0.$$



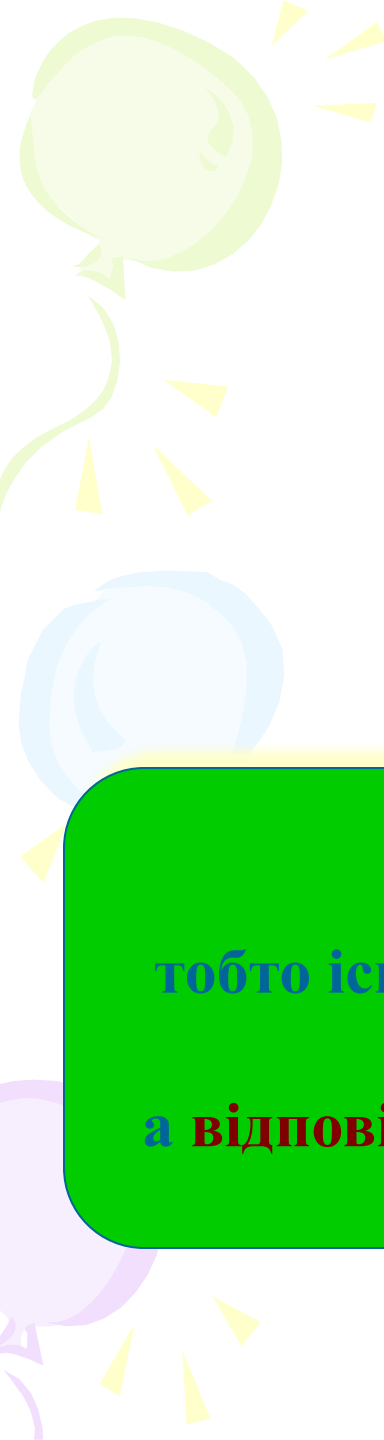
Альтернативною до неї є H_1 :
*значення хоча б одного параметра
моделі відмінне від нуля*, тобто
хоча б один із факторів впливає
на змінювання залежної змінної.

Для перевірки цих гіпотез застосовують *F-критерій Фішера* з *$n-m-1$* ступенями свободи.

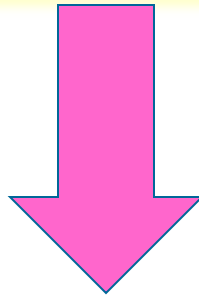
$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}$$

яке порівнюють з табличним значенням розподілу Фішера при заданому рівні значущості *α* .

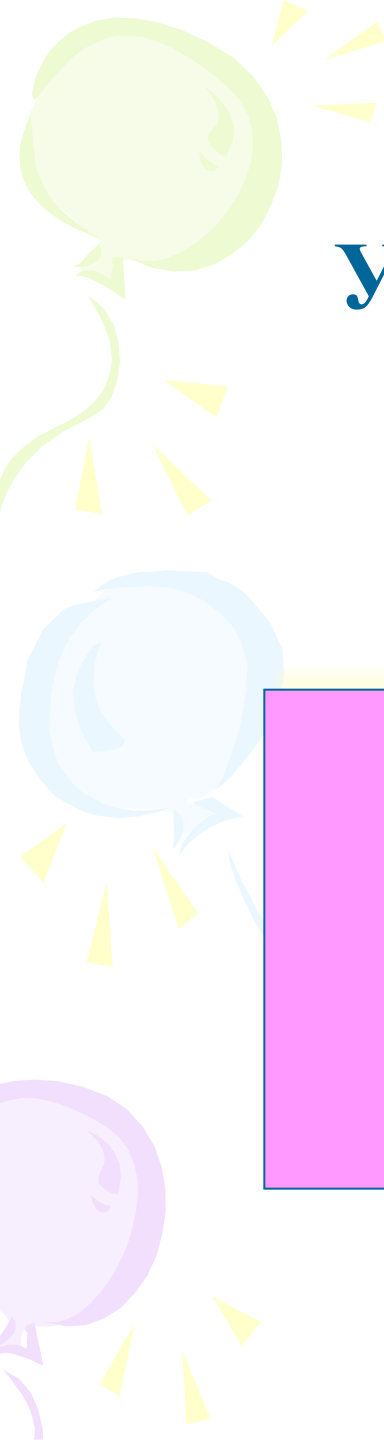
(Як правило, *$\alpha = 0,05$* або *$\alpha = 0,01$*).



Якщо $F_{табл} < F_{експ}$



**нульова гіпотеза відхиляється,
тобто існує такий коефіцієнт у регресійному рівнянні,
який суттєво відрізняється від нуля,
а відповідний фактор впливає на досліджувану змінну.**



У випадку парної регресії цей критерій розраховується за формулою:

$$F = \frac{\sum_{i=1}^n (\hat{y} - \bar{y})^2}{1} : \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n - 2}$$



Коефіцієнт кореляції, як вибіркова характеристика, перевіряється на значущість за допомогою ***t-критерію Стьюдента.***

Фактичне значення *t* статистики обчислюється за формулою

$$t_{\text{експ}} = \frac{R \sqrt{n - m - 1}}{\sqrt{1 - R^2}}$$



$t_{\text{експ}}$ порівнюється з табличним значенням t -розподілу з $n - m - 1$ ступенями свободи, та при заданому рівні значущості $\alpha/2$

Якщо $|t_{\text{експ}}| > t_{\text{табл}}$



можна зробити висновок, що коефіцієнт кореляції **достовірний** (значущий), а зв'язок між залежною змінною та всіма незалежними факторами **суттєвий**.

Можна визначити *стандартні похибки* оцінок параметрів моделі з урахуванням дисперсії залишків:

$$S_{\hat{a}_j} = \sqrt{\sigma_u^2 \cdot c_{jj}}$$

де σ_u^2 - дисперсія залишків, обчислюється за формулою:

$$\sigma_u^2 = \frac{\sum_{i=1}^n u_i^2}{n - m}$$

c_{jj} – відповідний діагональний елемент матриці похибок C (матриця, обернена до матриці коефіцієнтів системи нормальних рівнянь)

$$C = A^{-1} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1}$$

Статистичну значущість кожного параметра моделі можна перевірити за допомогою ***t*-критерію**.

При цьому нульова гіпотеза буде:

$H_0: a_j = 0$, альтернативна **$H_1: a_j \neq 0$** Експериментальне значення

t-статистики для кожного параметра моделі обчислюється за формулою:

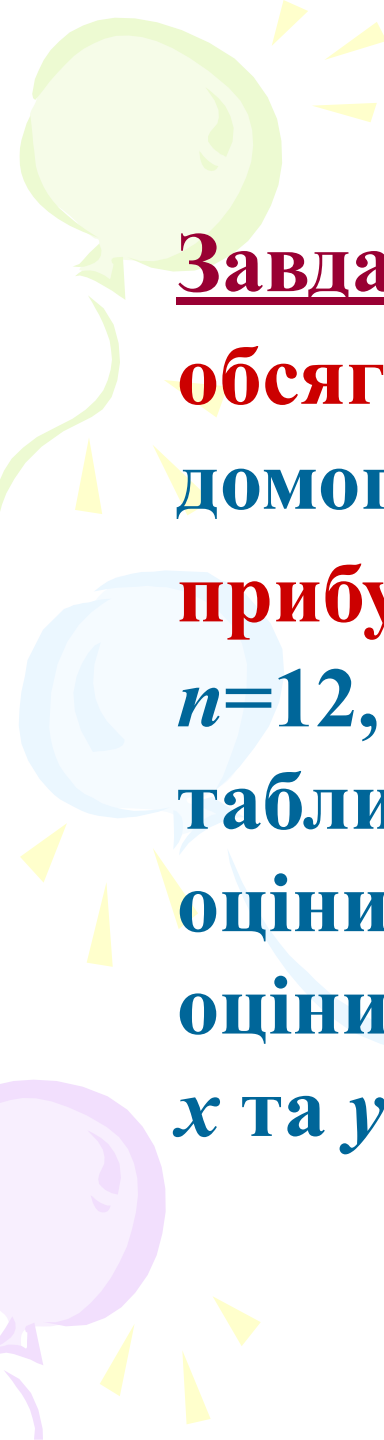
$$t_j = \frac{a_j}{S_{a_j}}$$

Довірчі інтервали для кожного параметра a_j обчислюються на основі його стандартної похибки та критерію Стюдента:

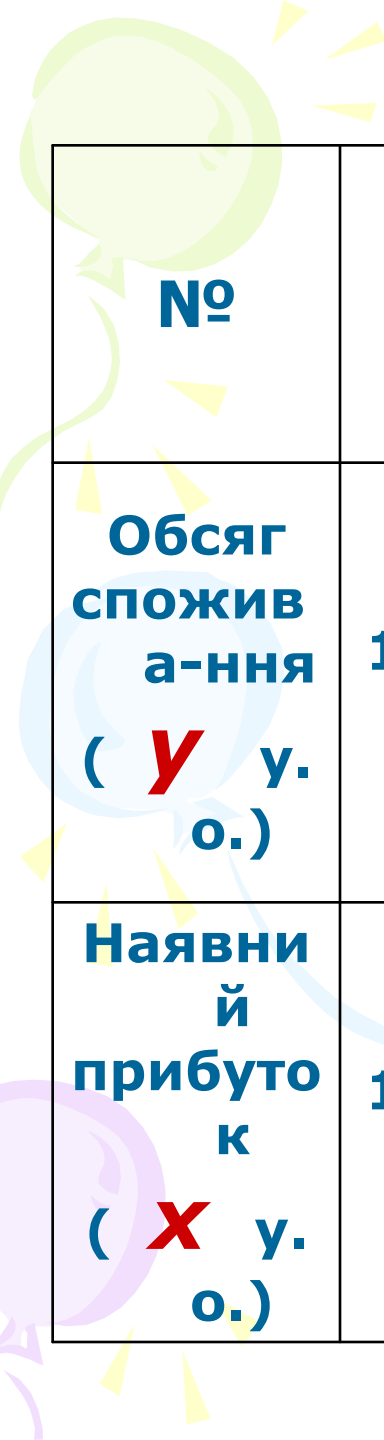
$$\left(a_j - t_{\text{табл}} \sqrt{\sigma_u^2 \cdot c_{jj}} \quad ; \quad a_j + t_{\text{табл}} \sqrt{\sigma_u^2 \cdot c_{jj}} \right)$$

Табличне значення $t_{\text{табл.}}$, як і раніше, має $n-m-1$ ступенів свободи і рівень значущості $\alpha/2$

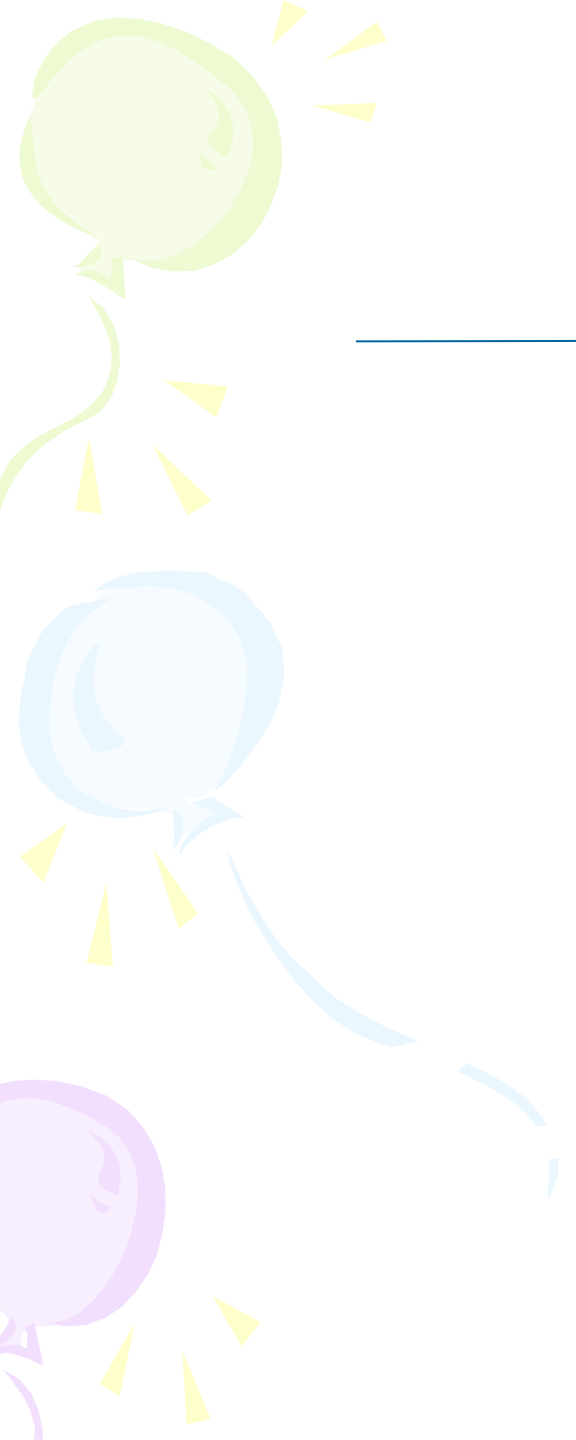
$$t_{\text{табл}} = t_{\alpha/2}(n - m - 1)$$



Завдання: Зробити аналіз залежності
обсягу споживання y (у.о.)
домогосподарства від наявного
прибутку x (у.о.) за вибіркою обсягом
 $n=12$, результати якої наведено у
таблиці. Визначити вид залежності,
оцінити параметри рівняння регресії,
оцінити силу лінійної залежності між
 x та y .



№	1	2	3	4	5	6	7	8	9	10	11	12
Обсяг спожива- ння (у. о.)	107	109	110	113	120	122	123	128	136	140	145	150
Найви- щий прибуток (у. о.)	102	105	108	110	115	117	119	125	132	130	141	144



$y = 3,423 + 0,936 x$

Regression Summary for Dependent Variable: y (labor2.sta)						
R= ,99160670 R_ = ,98328384 Adjusted R_ = ,98161222						
F(1, 10)=588,22 p<,00000 Std.Error of estimate: 1,8775						
N=12	Beta	Std.Err. of Beta	B	Std.Err. of B	t(10)	p-level
Intercept			3,422610	4,864432	0,70360	0,497738
x	0,991607	0,040885	0,936080	0,038596	24,25332	0,000000

коефіцієнт кореляції

$R=0,9916067$.

Значення критерію Стюдента

**Коефіцієнт
детермінації .**

$R^2 = 0,98328384$

**Значення
критерію Фішера**