

Аналіз даних та статистичне виведення

Тиждень

3



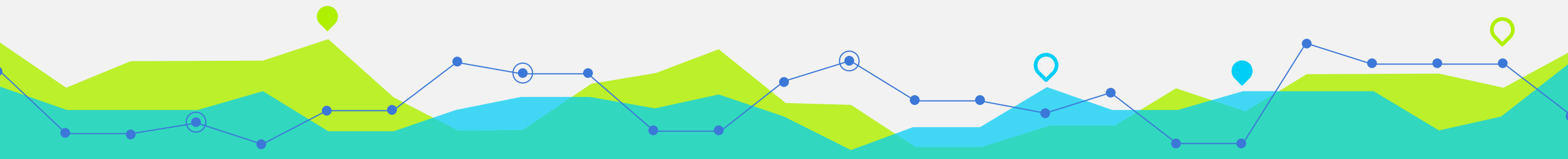
теорія ймовірності

ОСНОВНІ ПОНЯТТЯ

Випробування (експеримент) – сукупність умов, за яких спостерігається певне явище чи результат

Подія – факт, який а результаті експерименту може відбутись чи не відбутись

Ймовірність – чисельна міра впевненості в появі даної події внаслідок нового випробування



ймовірність

$P(A)$ – ймовірність настання події A

$$0 \leq P(A) \leq 1$$

Принцип практичної неможливості малоїмовірних подій: якщо випадкова подія має дуже малу ймовірність, то практично можна вважати, що в одиничному випробуванні подія не наступить.

Ймовірність

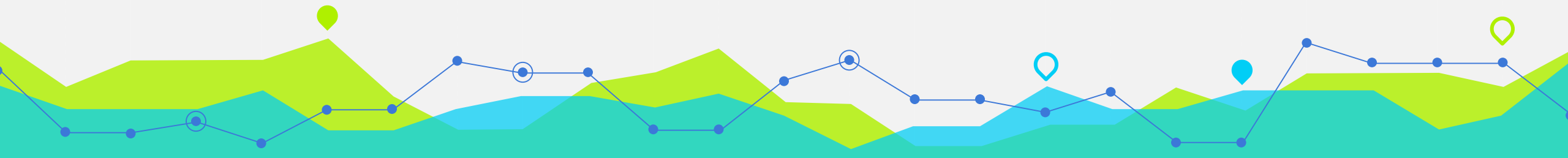
Частотна інтерпретація

$$P(A) = m/n$$

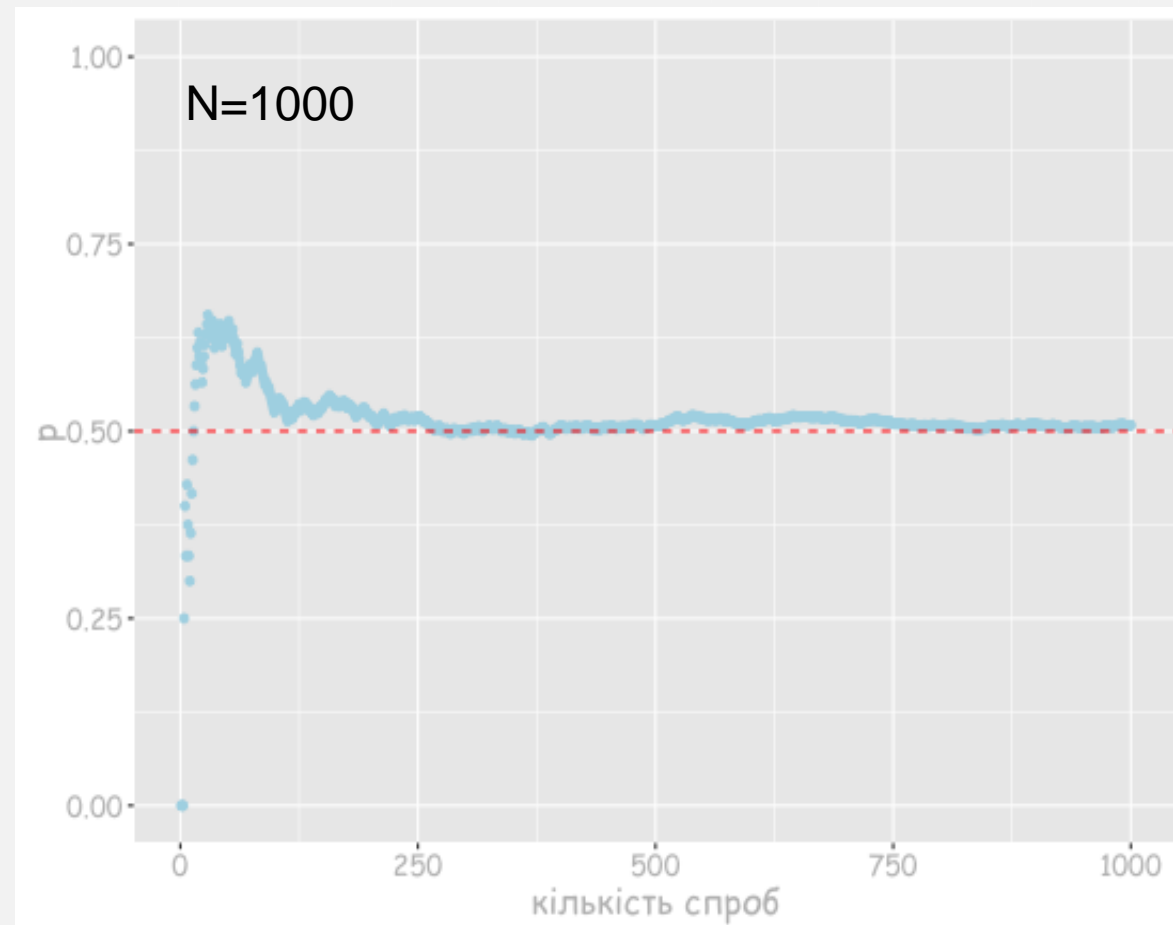
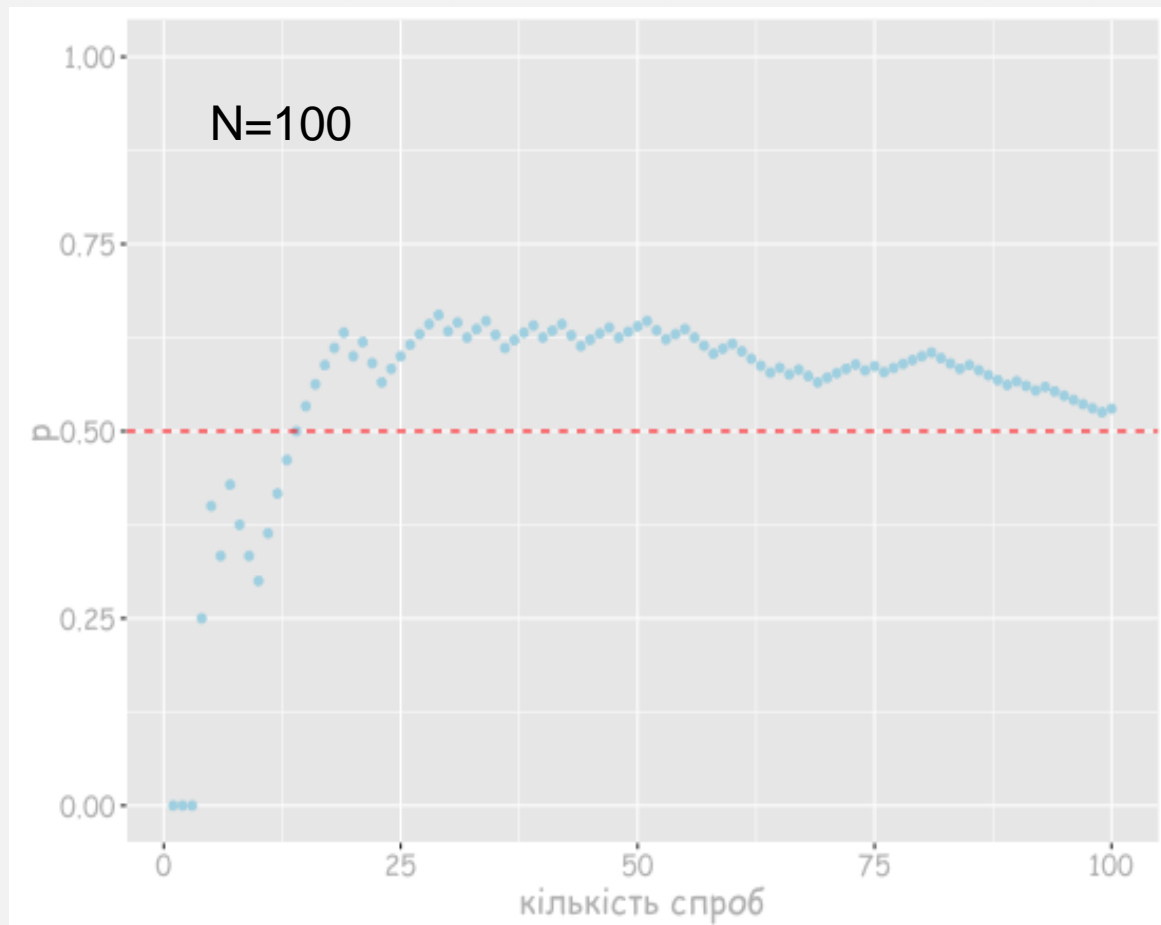
$P(A)$ – ймовірність події A ,
 m – кількість випадків, що
сприяють настанню події A ,
 n – загальна кількість
випадків. $P(A)$

Байєсівська інтерпретація

оцінка ймовірності через
суб'єктивну ступінь
впевненості в настанні події
через спостереження



закон великих чисел





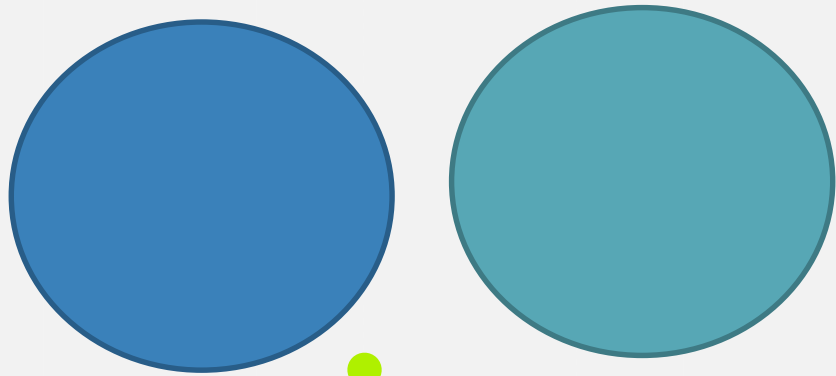
ймовірність кількох
подій

Ймовірність кількох подій

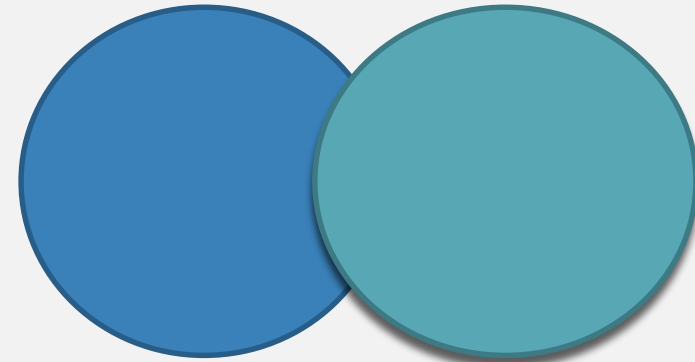
Несумісні події не можуть відбуватись одночасно.

Сумісні події – можуть відбутись одночасно

Взаємовиключні події – коли настає одна або друга подія



$$P(A \text{ and } B) = 0$$



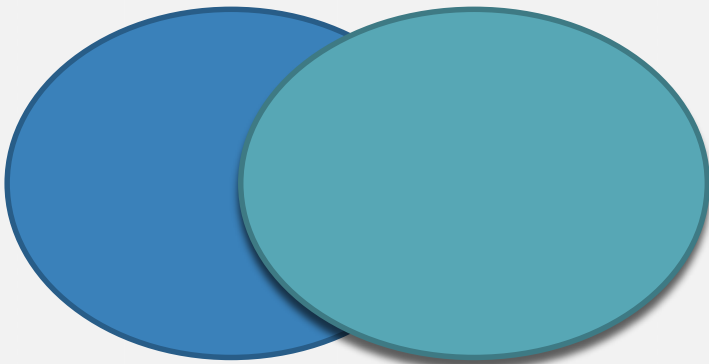
$$P(A \text{ and } B) \neq 0$$

Було опитано 77882 людини з 57 країн світу. 36.2% погоджуються з твердженням “Чоловіки повинні мати більше прав ніж жінки”. 13.8% мають університетську освіту.

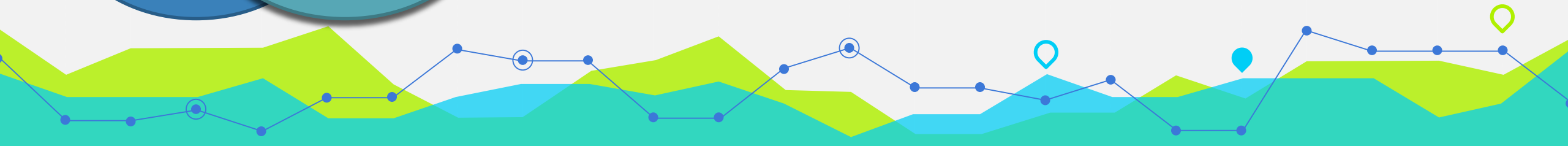
3.6% належать до обох категорій.

Яка ймовірність, що випадковим чином обрана людина з вищою освітою або погоджується з твердженням “Чоловіки повинні мати більше прав ніж жінки”?

$$P(A \text{ or } B) = p(A) + p(B) - p(A \text{ and } B)$$



Source: wordvaluesurvey.org



Ймовірність кількох подій

Дві події є **несумісними** якщо вони не можуть відбуватися одночасно.

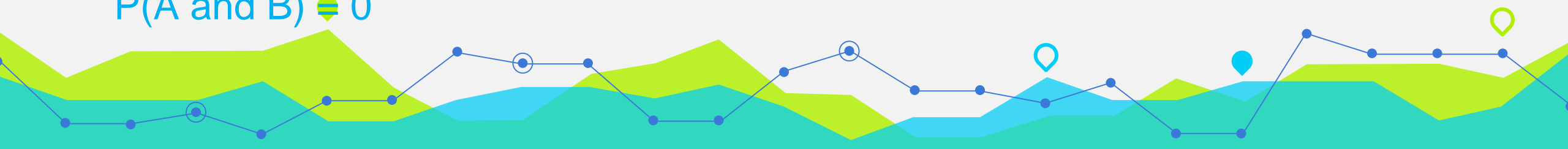
$$P(A \text{ and } B) = 0$$

Дві події є **незалежними**, якщо знання про настання однієї з них не дає можливості оцінити ймовірність настання іншої.

$$P(A | B) = P(A)$$

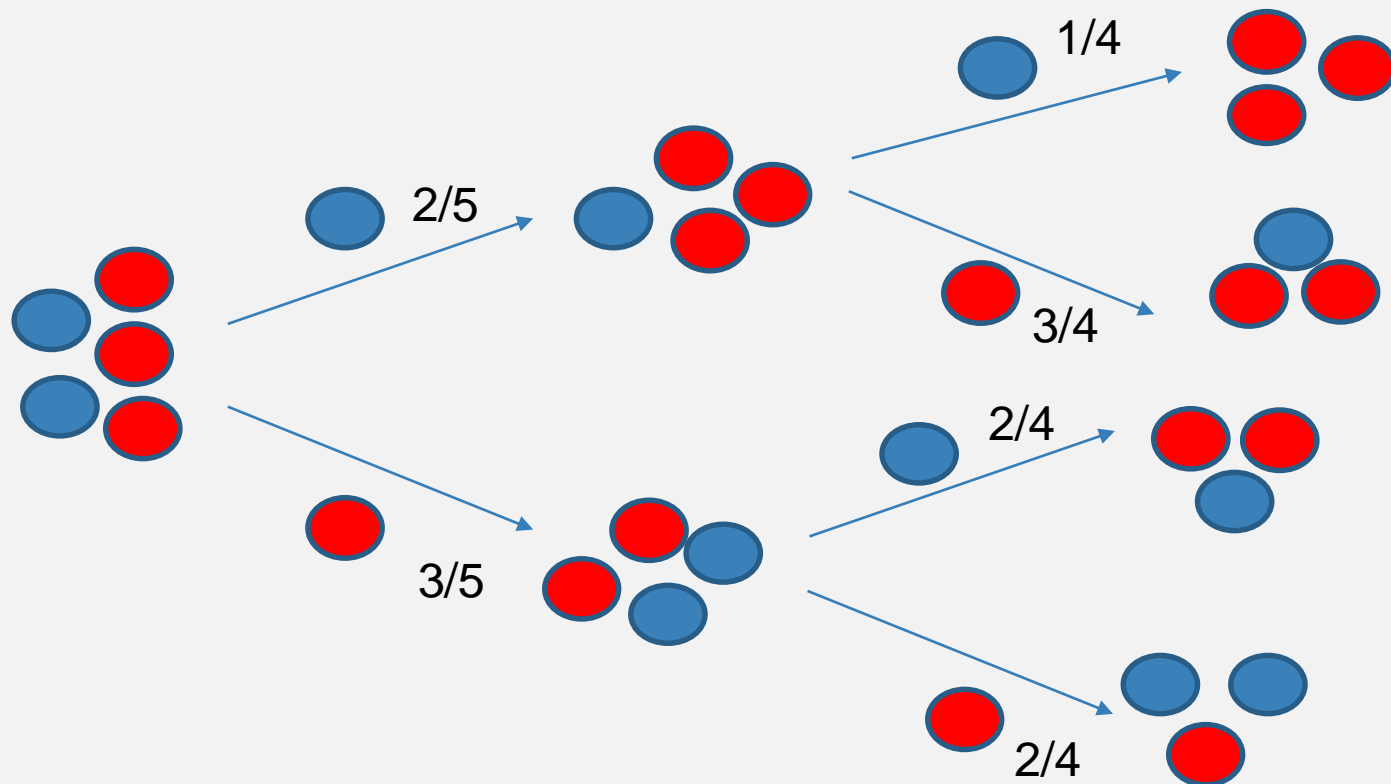
Дві події є **залежними**, якщо знання про настання однієї з них змінює ймовірність настання іншої.

$$P(A | B) = P(A) P(B | A)$$



умовні ймовірності

Умовна ймовірність – ймовірність настання події за умови настання іншої події



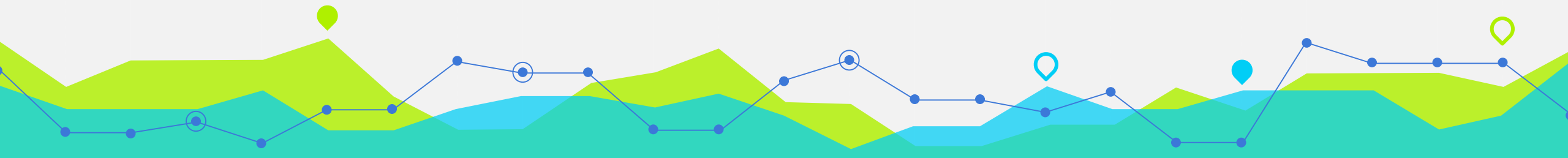
Ймовірність кількох подій

Об'єднання взаємовиключних подій: $P(A \text{ or } B) = P(A) + P(B)$

Об'єднання сумісних подій: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Переріз незалежних подій: $P(A \text{ and } B) = P(A) \times P(B)$

Переріз залежних подій: $P(A \text{ and } B) = P(A) \times P(B | A)$

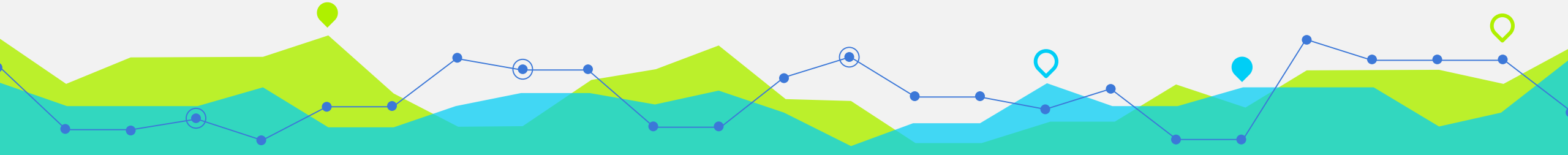
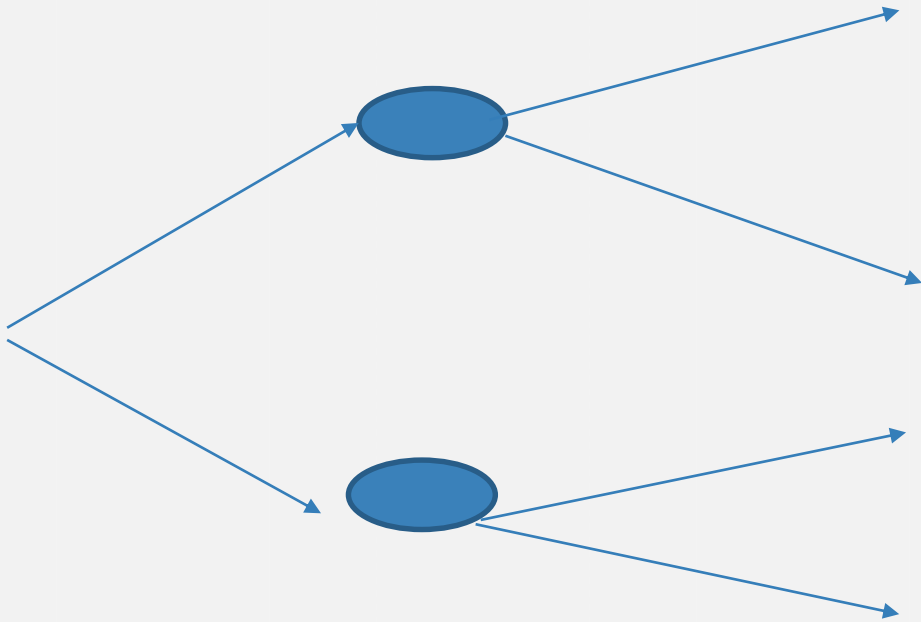


У екзаменаційному тесті є п'ять питань, кожне з яких має чотири варіанти відповіді. Яка ймовірність правильно відповісти на всі п'ять, якщо ви не готувалися до цього екзамену?



дерева прийняття рішень

У вашій електронній скриньці 100 повідомлень. 60 з них це спам. З цих 60 листів 35 мають в тексті слово “безкоштовно”. З решти 40 листів 3 мають в тексті слово “безкоштовно”. Якщо в повідомленні є безкоштовно, яка ймовірність що це спам?





теорема Байеса

теорема Байеса

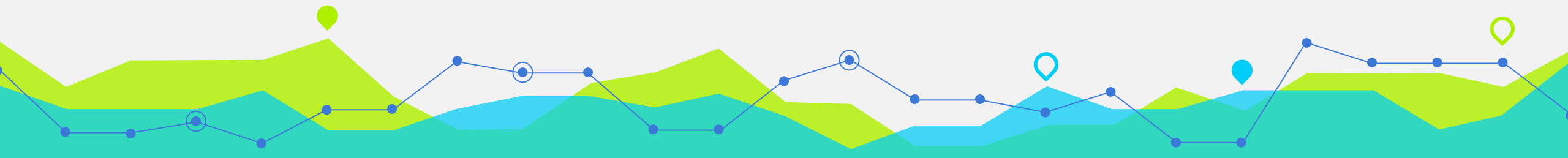
$$P(A | B) = \frac{P(B | A) * P(A)}{P(B)}$$

$P(A)$ та $P(B)$ ймовірності настання подій A та B

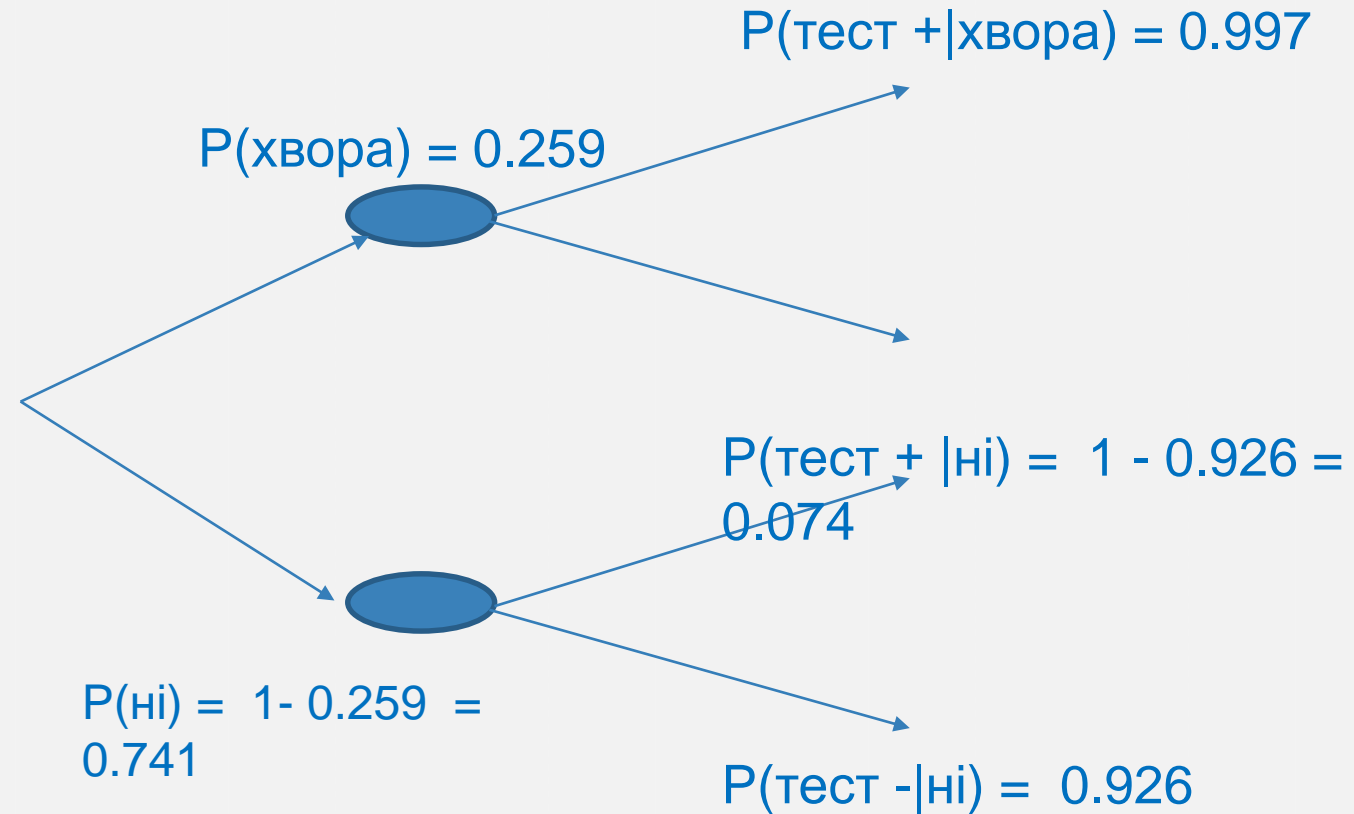
$P(A|B)$ ймовірність настання події A , якщо відбулася подія B

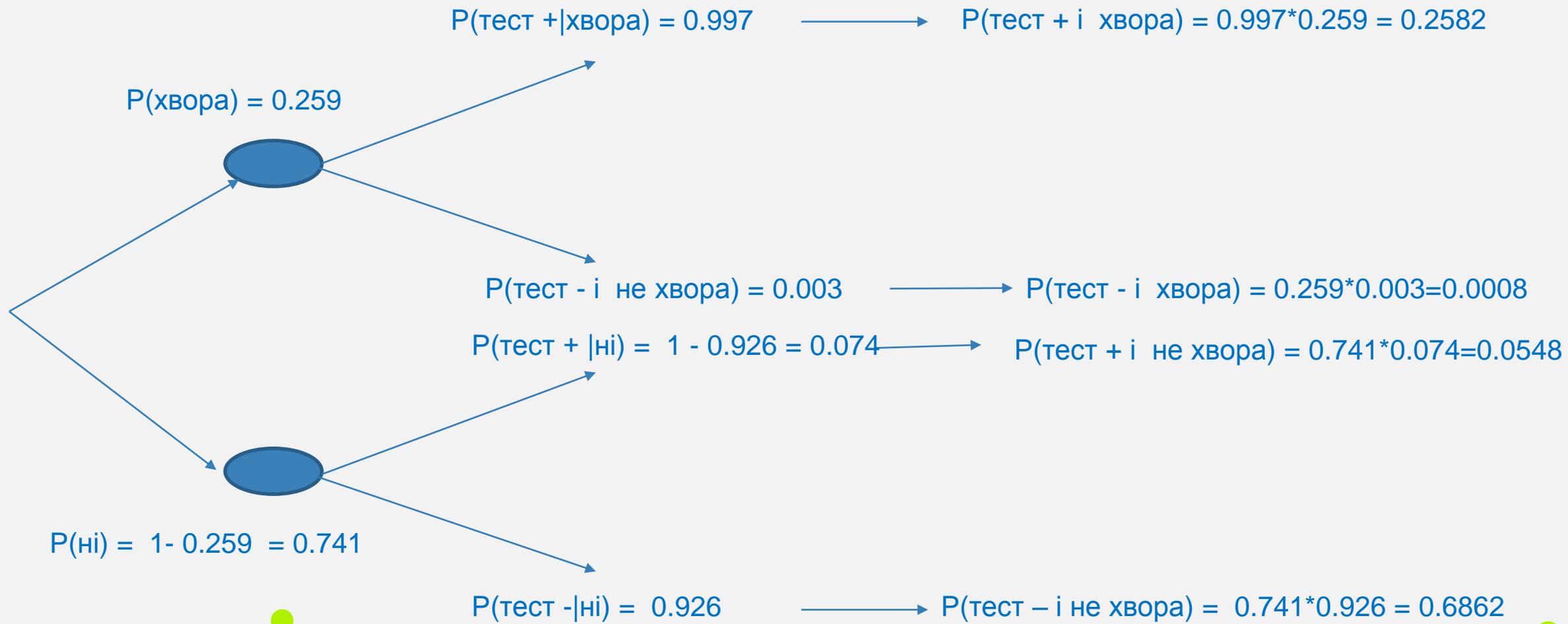
$P(B|A)$ ймовірність настання події B , якщо відбулася подія A

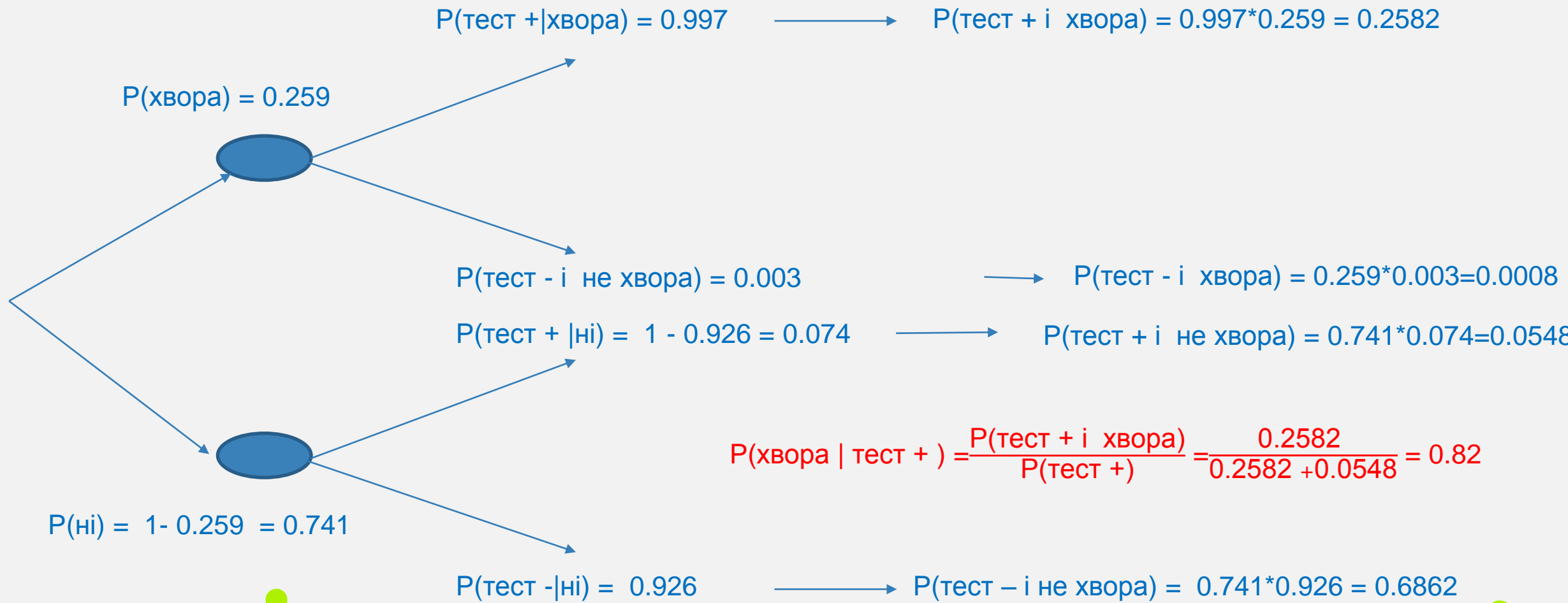
$$P(A | B) = \frac{P(B | A) * P(A)}{P(B | A) * P(A) + P(B|\neg A) * P(\neg A)}$$



В 2009 році в найвищий відсоток захворюваності на ВІЛ/СНІД було зафіксовано в Свaziленді і становить 25.9% Тест ELISA один з найкращих та найточніших тестів. Для тих, хто хворий на СНІД тест має точність 99.7%, для тих хто не хворий 92.6%. Якщо за результатами тесту людина ВІЛ інфікована, яка ймовірність що вона дійсно хвора?







Марі виходить заміж завтра. Церемонія буде відбуватися надворі. Марі живе у пустельній місцевості. Останні роки дощ йшов в середньому 5 днів в році. На жаль, синоптик передбачив дощ на завтра. Відомо, що метеоролог правильно прогнозує дощ у 90% випадків. Коли дощу не буде, він помиляється (тобто неправильно прогнозує дощ) у 10% випадків. Яка ймовірність того, що в день весілля Марі буде дощ?

Подія A_1 . Під час весілля Марі буде дощ

Подія A_2 . Під час весілля Марі дощу не буде

Подія В. Метеоролог пророкує дощ.



$P(A_1) = 5/365 = 0.0136985$ [Протягом року в середньому 5 дощових днів]

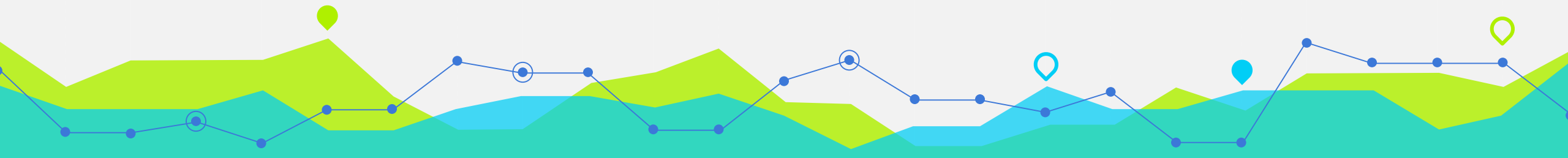
$P(A_2) = 360/365 = 0.9863014$ [Решта днів у році без осадів]

$P(B | A_1) = 0.9$ [Синоптик правильно передбачає дощ в 90% випадків]

$P(B | A_2) = 0.1$ [Синоптик передбачає дощ і помиляється у 10% випадків]

$$P(A_1 | B) = \frac{P(A_1) * P(B | A_1)}{P(A_1) * P(B | A_1) + P(A_2) * P(B | A_2)}$$

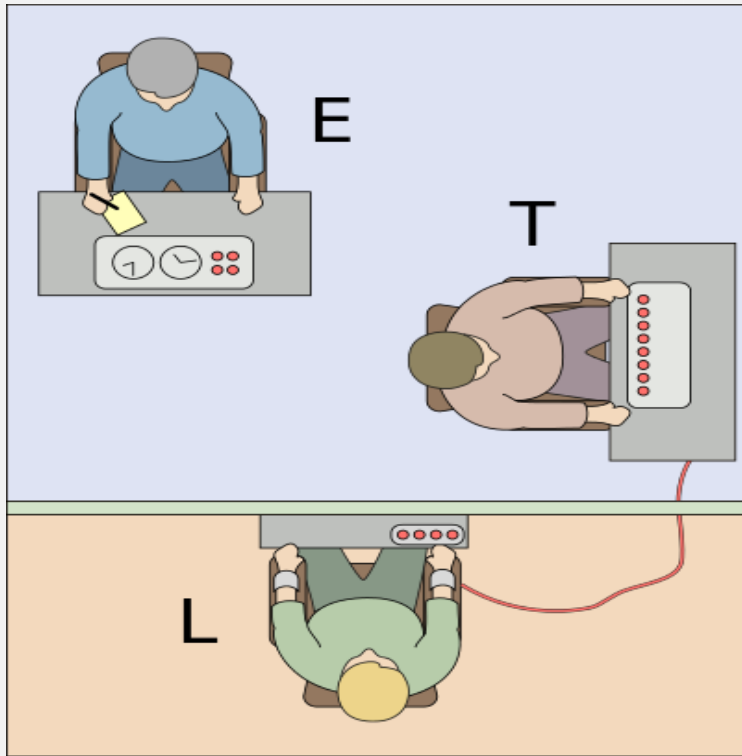
$$P(A_1 | B) = (0.014) * (0.9) / (0.014 * 0.9 + 0.986 * 0.1) = 0.111$$





біноміальний розподіл

експеримент Мілгрема



- Кожна особа в експерименті Мілгрема як випробування
- Успіх – якщо відмовилась, невдача якщо погодилась
- Оскільки 35% відмовляється- ймовірність успіху 35%
- Коли кожне випробування має лише два можливих наслідки – його називають випробуванням Бернуллі

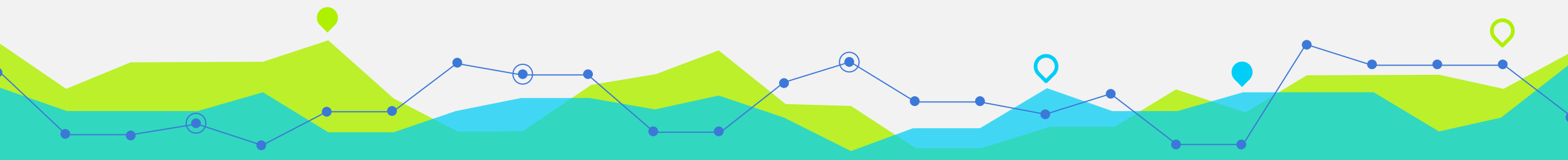
Якщо ми виберемо для експерименту трьох випадкових людей, яка ймовірність що один з них відмовиться ?

Антон #1: $0.35 * 0.65 * 0.65 = 0.149$

Богдан #2: $0.65 * 0.35 * 0.65 = 0.149$

Вікторія #3: $0.65 * 0.65 * 0.35 = 0.149$

} = 0.44



Скільки можливих сценаріїв
отримати 1 успіх у 4
випробуваннях?

Скільки можливих сценаріїв
отримати 2 успіхи в 9
випробуваннях?

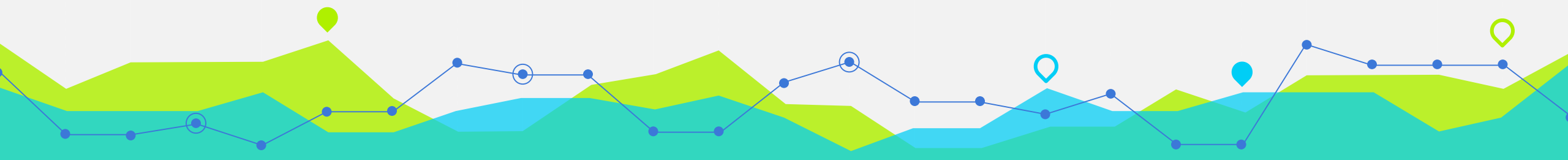


біноміальний розподіл

Якщо p репрезентує ймовірність успіху, $1-p$ – ймовірність невдачі, n – кількість незалежних випробувань, k – кількість успіхів.

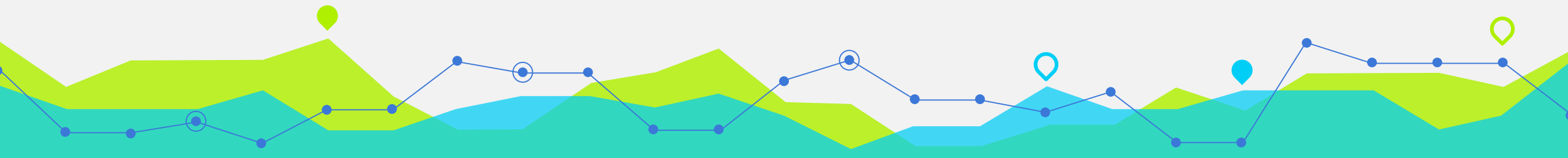
Ймовірність мати k успіхів в n незалежних випробуваннях Бернуллі з ймовірністю успіху p :

кількість сценаріїв \times $P(\text{одного сценарія})$

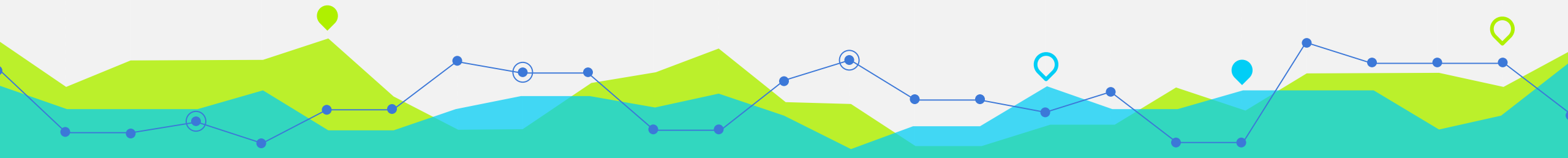
$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \times p^k (1-p)^{(n-k)}$$


УМОВИ

- Випробування незалежні
- Кількість випробувань n фіксована
- Кожен результат класифікується як успіх або невдача
- Ймовірність успіху p однакова для кожного випробування



Згідно опитування Gallup poll 2015 71% працівників віком 20-35 років незадоволені своїм місцем роботи. Яка ймовірність, що 4 з 10 випадковим чином обраних працівників незадоволені своїм місцем роботи?



Згідно опитування Gallup poll 2012 26.2% жителів США мають надмірну вагу. Яка ймовірність серед 20 випадковим чином обраних жителів отримати 5 з надлишковою вагою?



Математичне сподівання
біноміального розподілу:

$$\mu = np$$

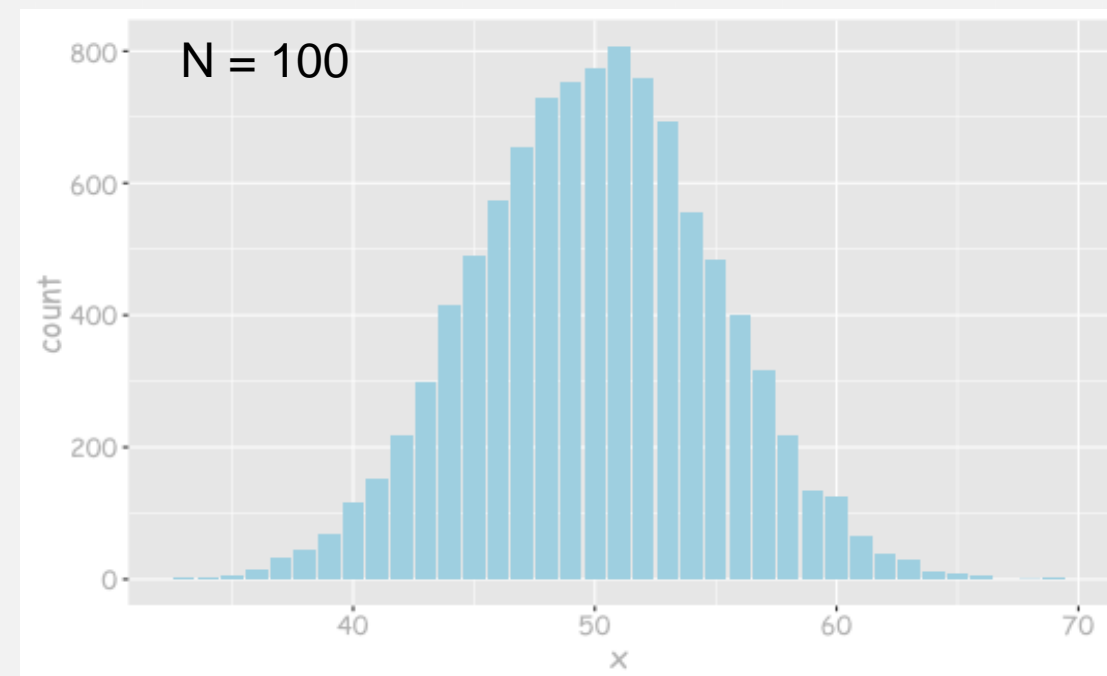
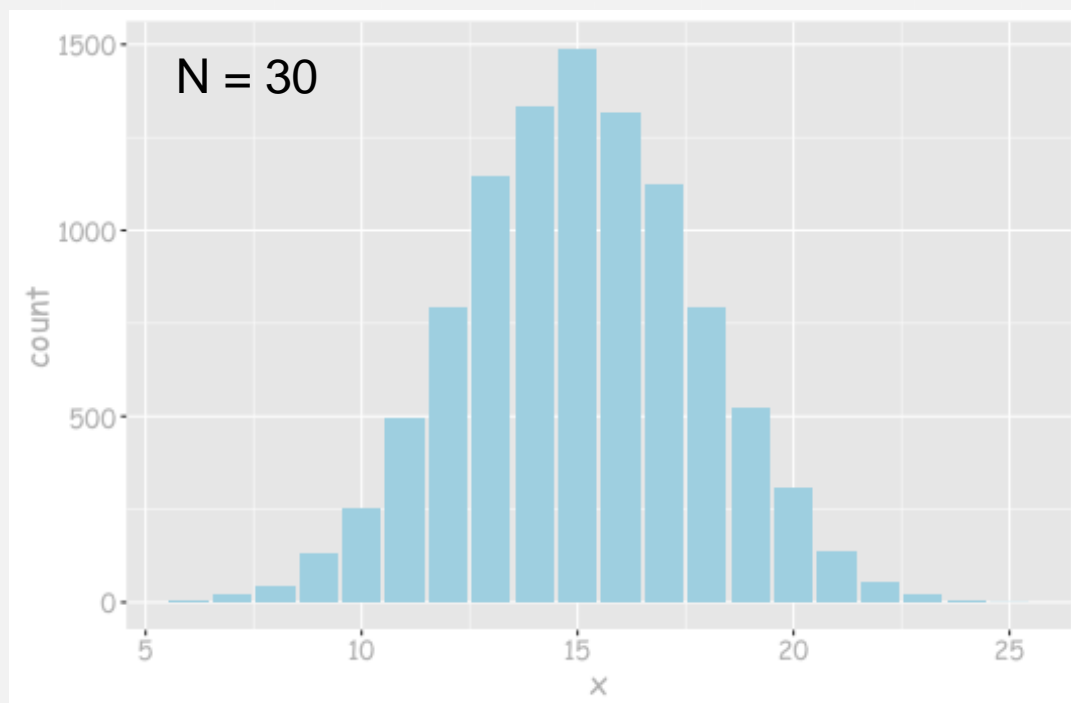
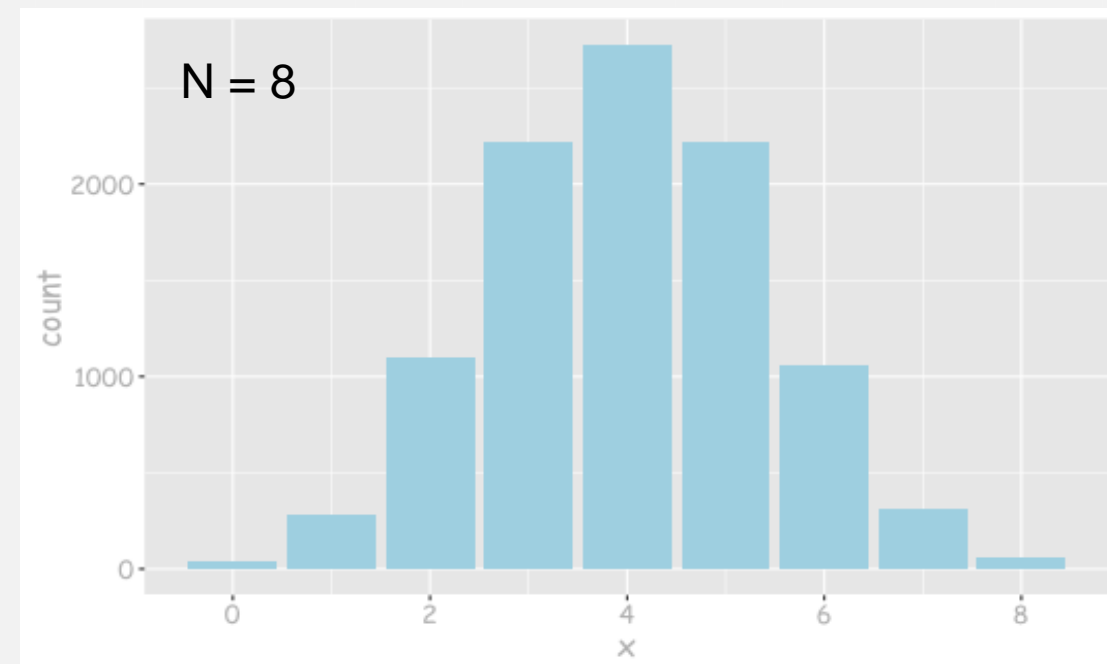
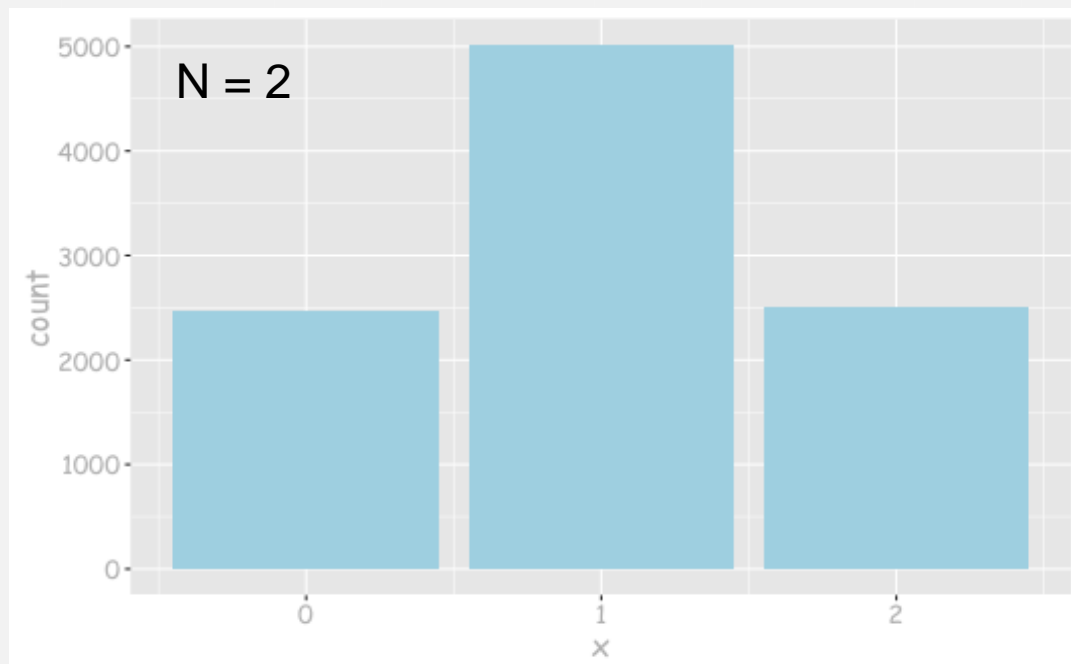
Середньоквадратичне
відхилення біноміального
розподілу:

$$\sigma = \sqrt{np(1 - p)}$$

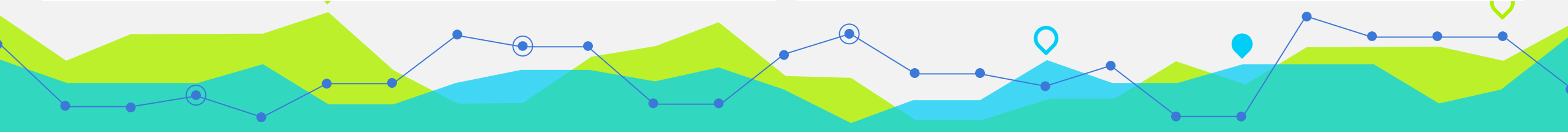




нормальний розподіл



нормальний розподіл



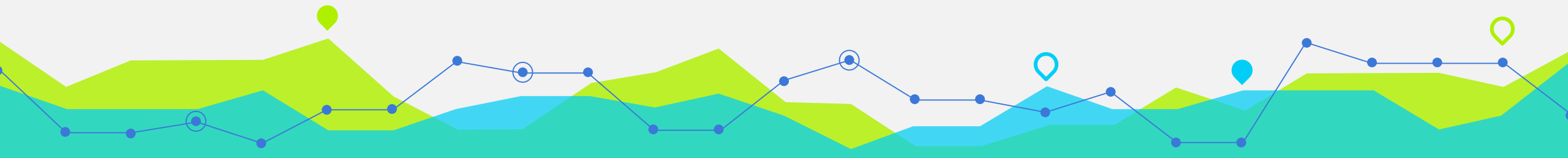
зв'язок між нормальним та біноміальним розподілами

Біноміальний розподіл, де очікується принаймі 15 успіхів та 15 невдач, поводить себе як нормальний розподіл.

$$np \geq 15, n(1-p) \geq 15$$

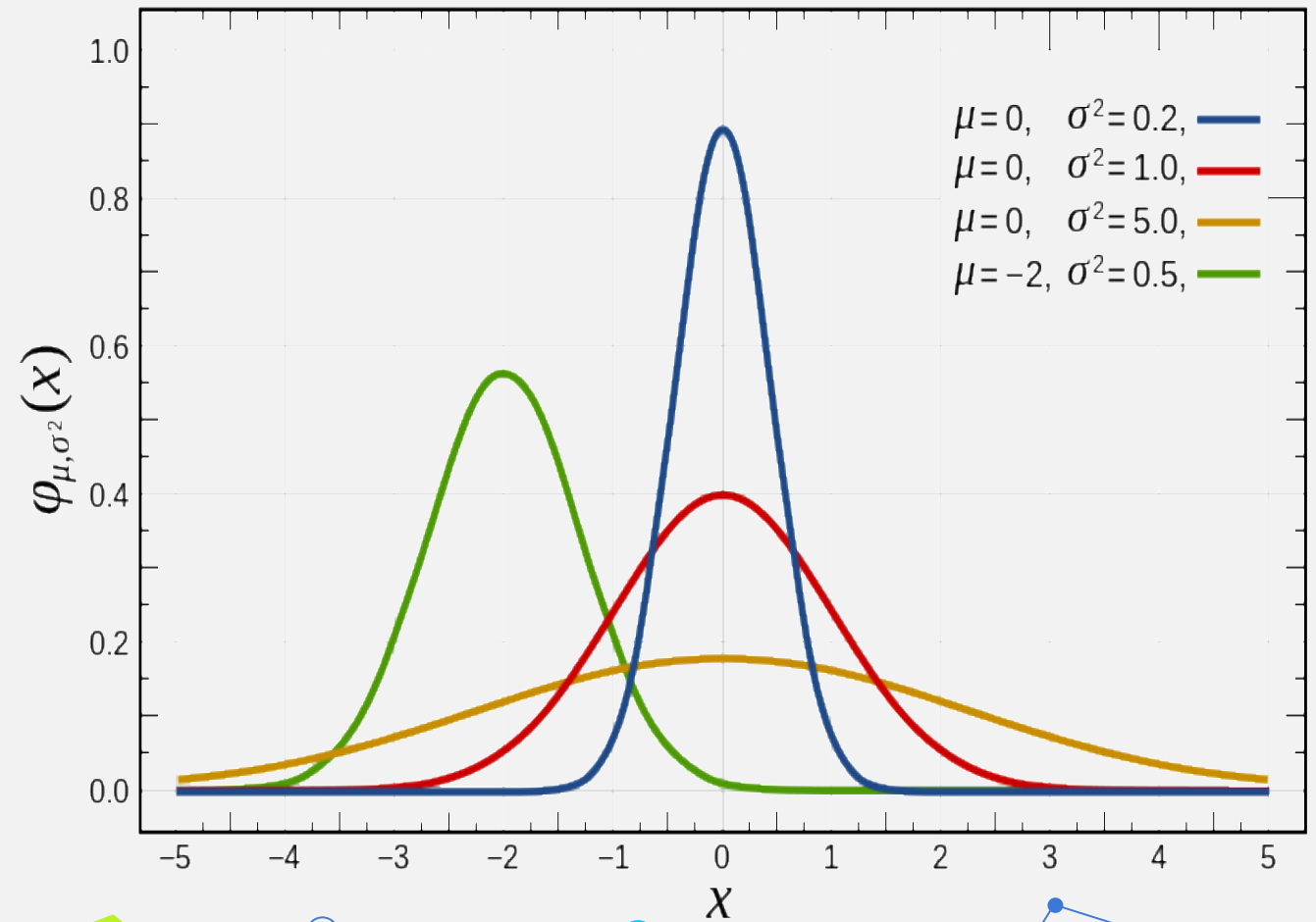
$$\text{Binomial}(n,p) \sim \text{Normal}(\mu, \sigma),$$

де $\mu=np, \sigma = \sqrt{np(1-p)}$



стандартний нормальний розподіл

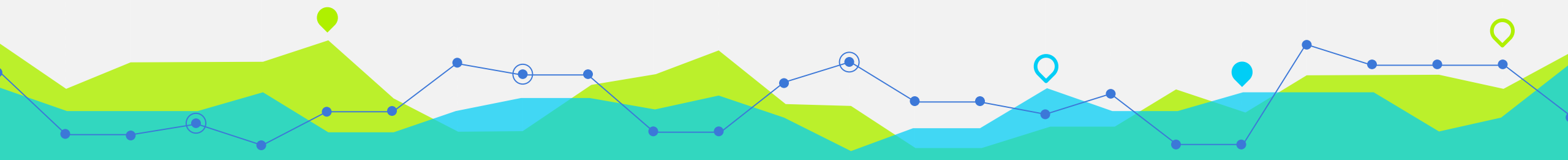
$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



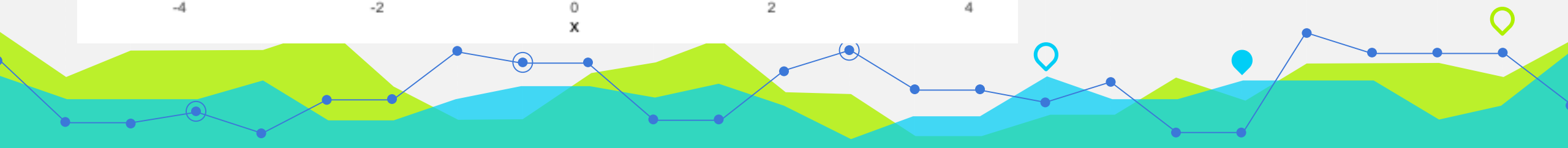
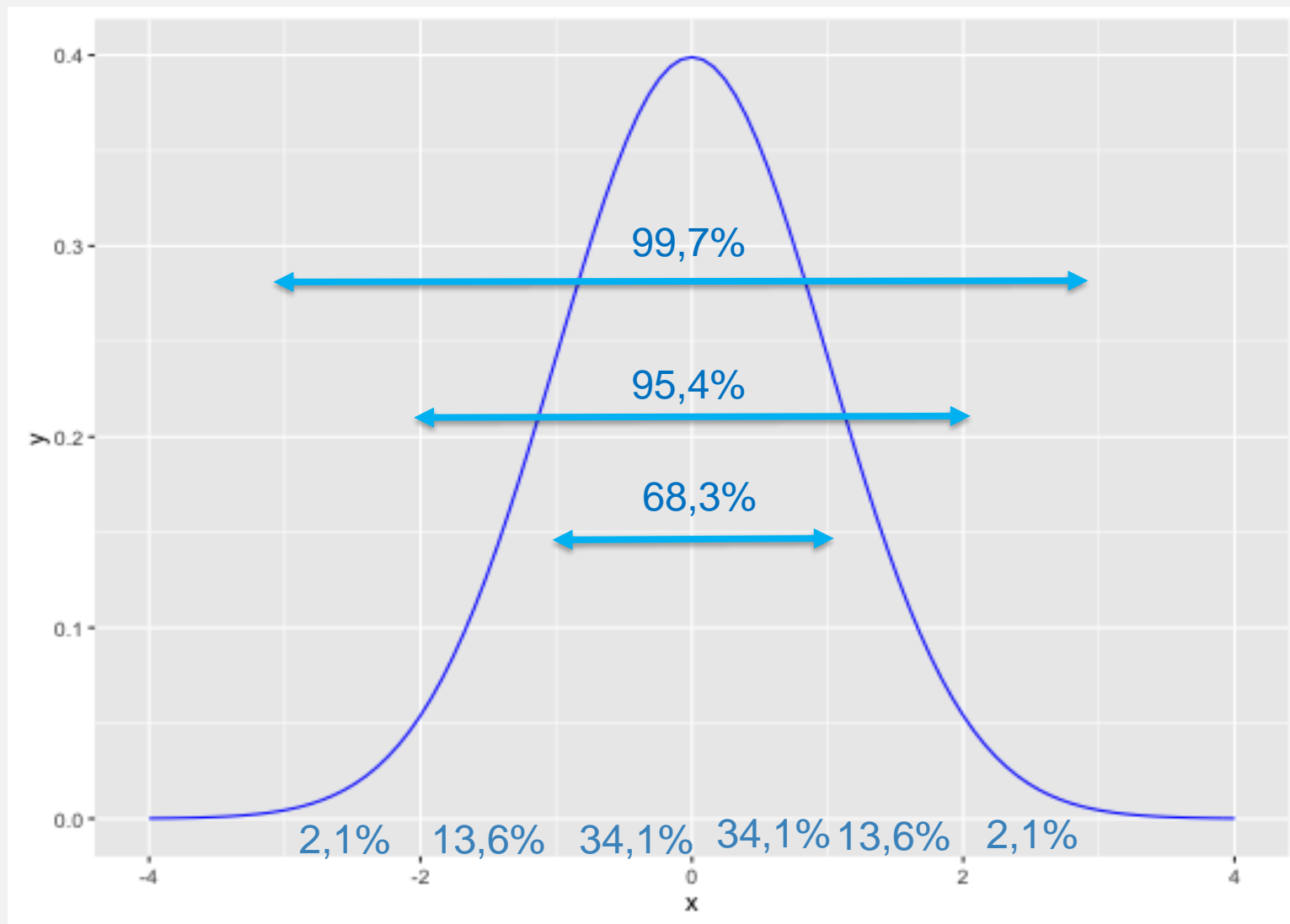
z-стандартизація

$$z = \frac{x - \mu}{\sigma}$$

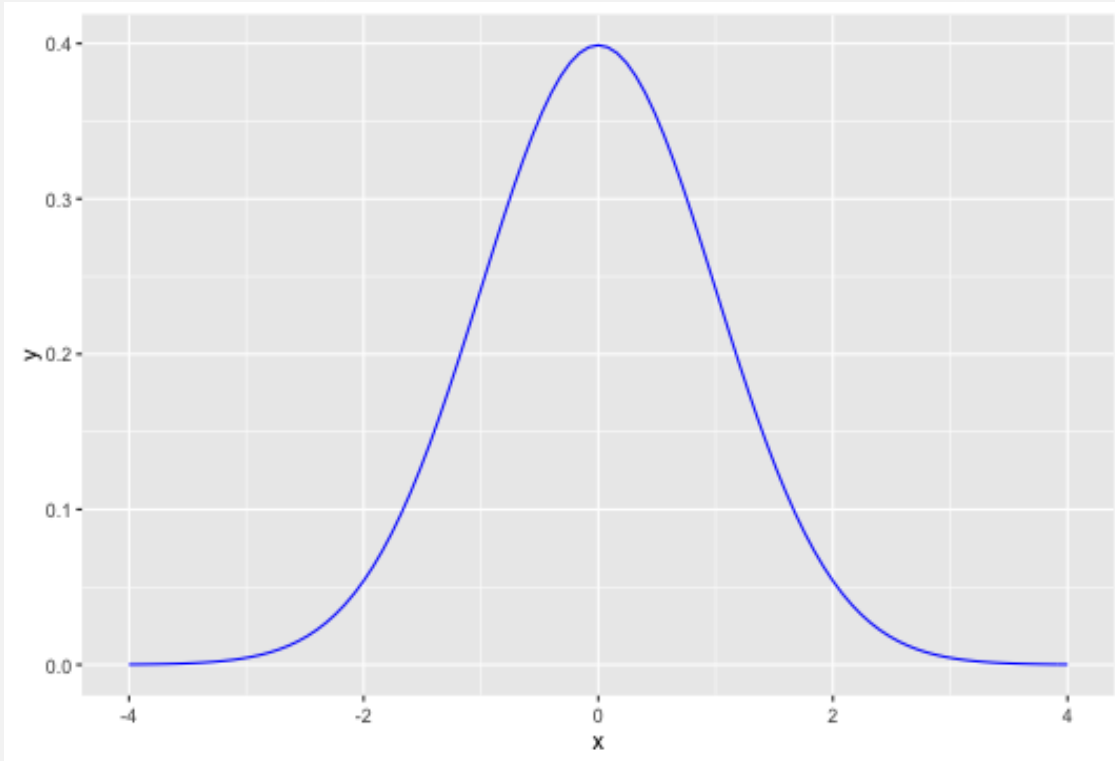
- Форма розподілу не змінюється
- Середнє значення стає нулем
- Середньоквадратичне відхилення стає одиницею



z-значення та ймовірність



як оцінити ймовірність отримати конкретне z-значення?



- 1)Робимо z-стандартизацію
- 2)Зображаємо наш розподіл
- 3)Визначаємо, який саме відрізок площі під кривою нас цікавить
- 4)Знаходимо значення в z-таблицях чи з допомогою функції `pnorm` в R

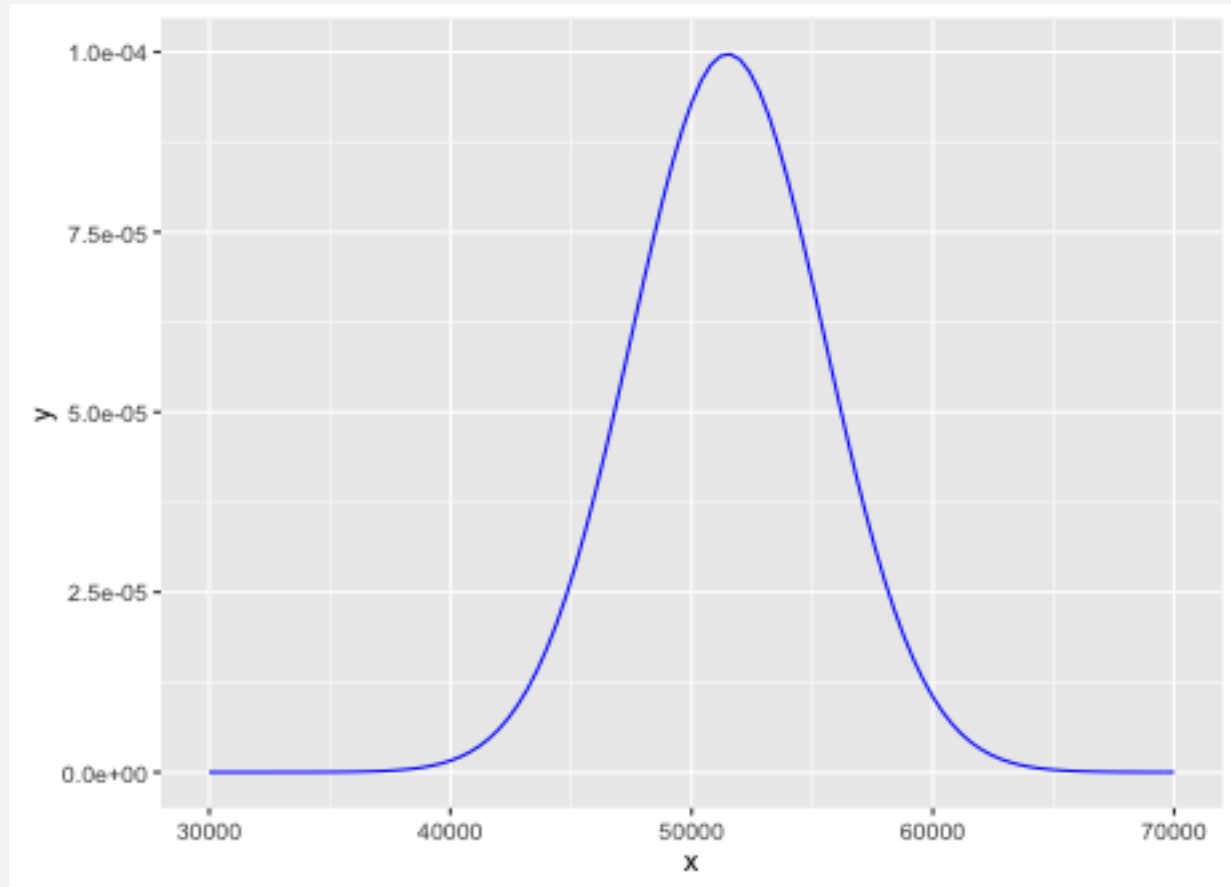
```
> pnorm(-1)
[1] 0.1586553
> pnorm(-1, 20, 10)
[1] 0.01786442
> pnorm(-1, 20, 10, lower.tail = FALSE)
[1] 0.9821356
```

https://gallery.shinyapps.io/dist_calc/

Ймовірність та z-значення

Виробник зимових шин декларує, що вони прослужать в середньому 51500 кілометрів та середньоквадратичне відхилення в 4000 кілометрів.

- Якщо ви придбаєте комплект таких шин, яка ймовірність, що вони послужитимуть принаймі 63 000 кілометрів?
- Який відсоток цих шин прослужить менше ніж 45000?
- Між 45000 і 55000?



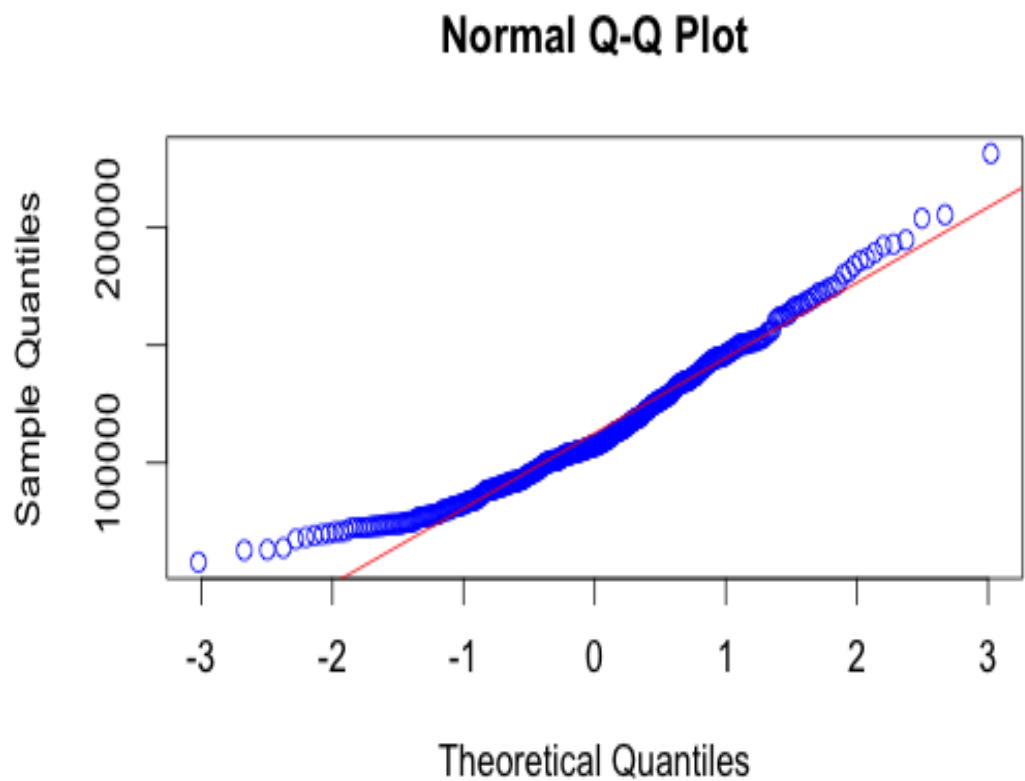
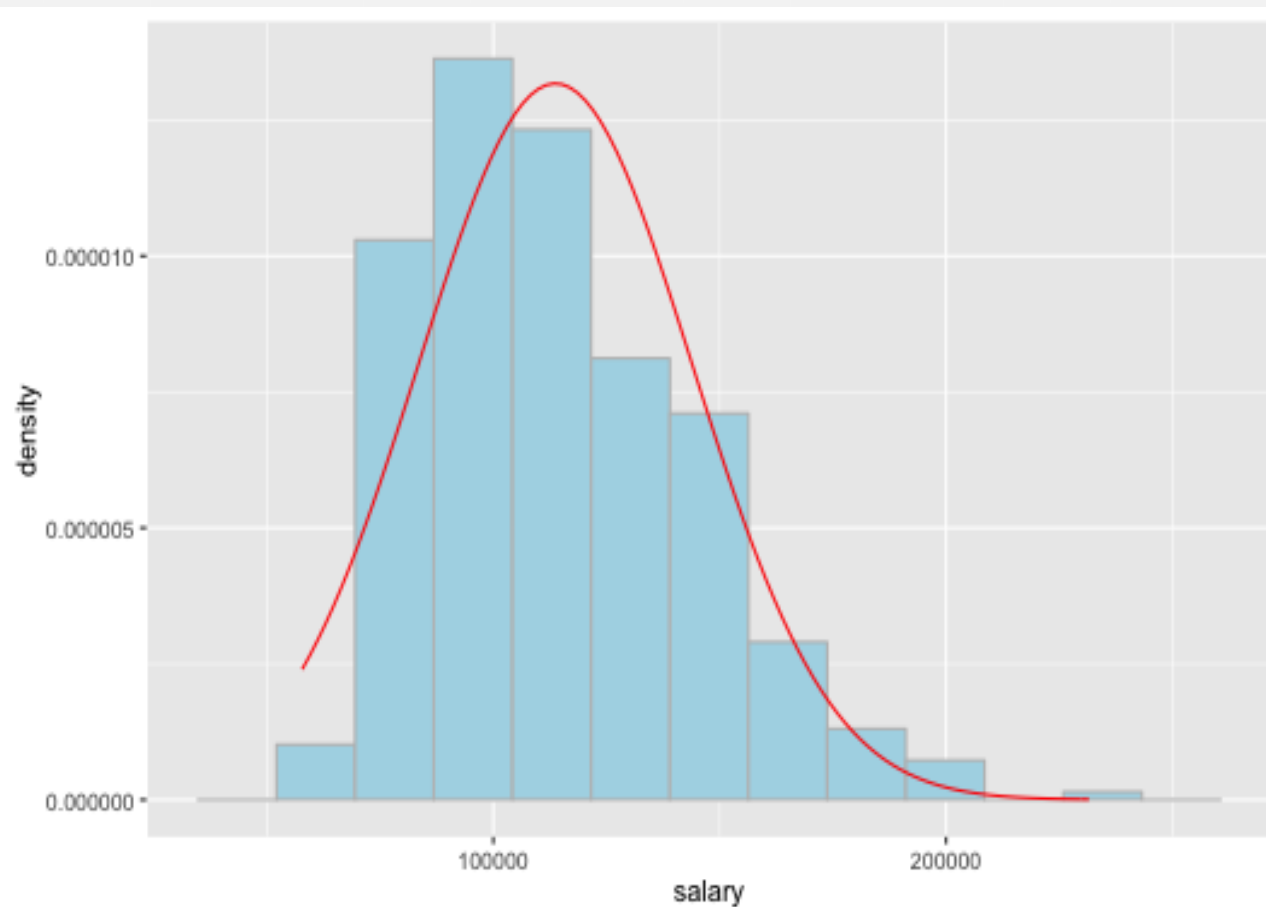
як перевірити чи розподіл є нормальним?

	X	rank	discipline	yrs.since.phd	yrs.service	sex	salary
1	1	Prof	B	19	18	Male	139750
2	2	Prof	B	20	16	Male	173200
3	3	AsstProf	B	4	3	Male	79750
4	4	Prof	B	45	39	Male	115000
5	5	Prof	B	40	41	Male	141500
6	6	AssocProf	B	6	6	Male	97000
7	7	Prof	B	30	23	Male	175000
8	8	Prof	B	45	45	Male	147765
9	9	Prof	B	21	20	Male	119250
10	10	Prof	B	18	18	Female	129000
11	11	AssocProf	B	12	8	Male	119800
12	12	AsstProf	B	7	2	Male	79800

<https://vincentarelbundock.github.io/Rdatasets/doc/car/Salaries.html>

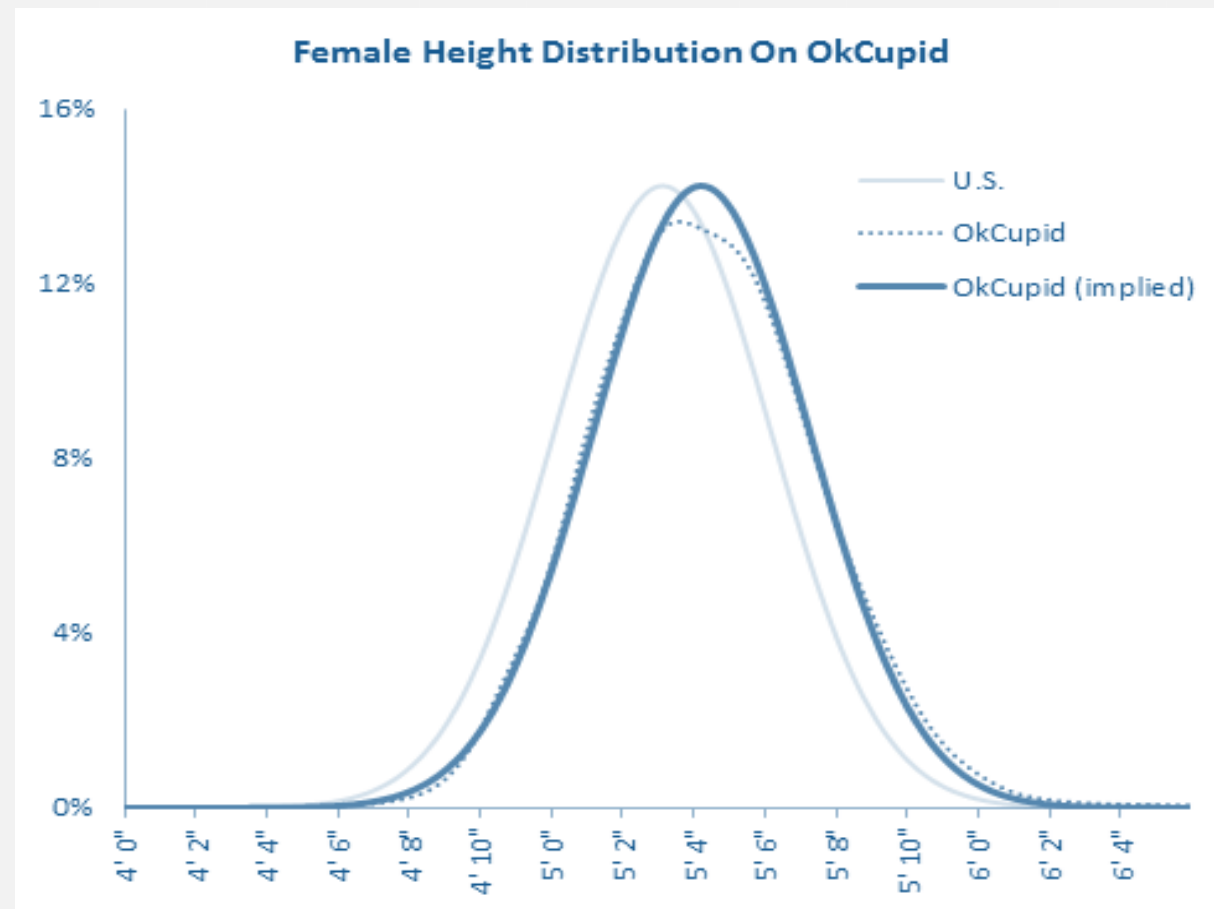
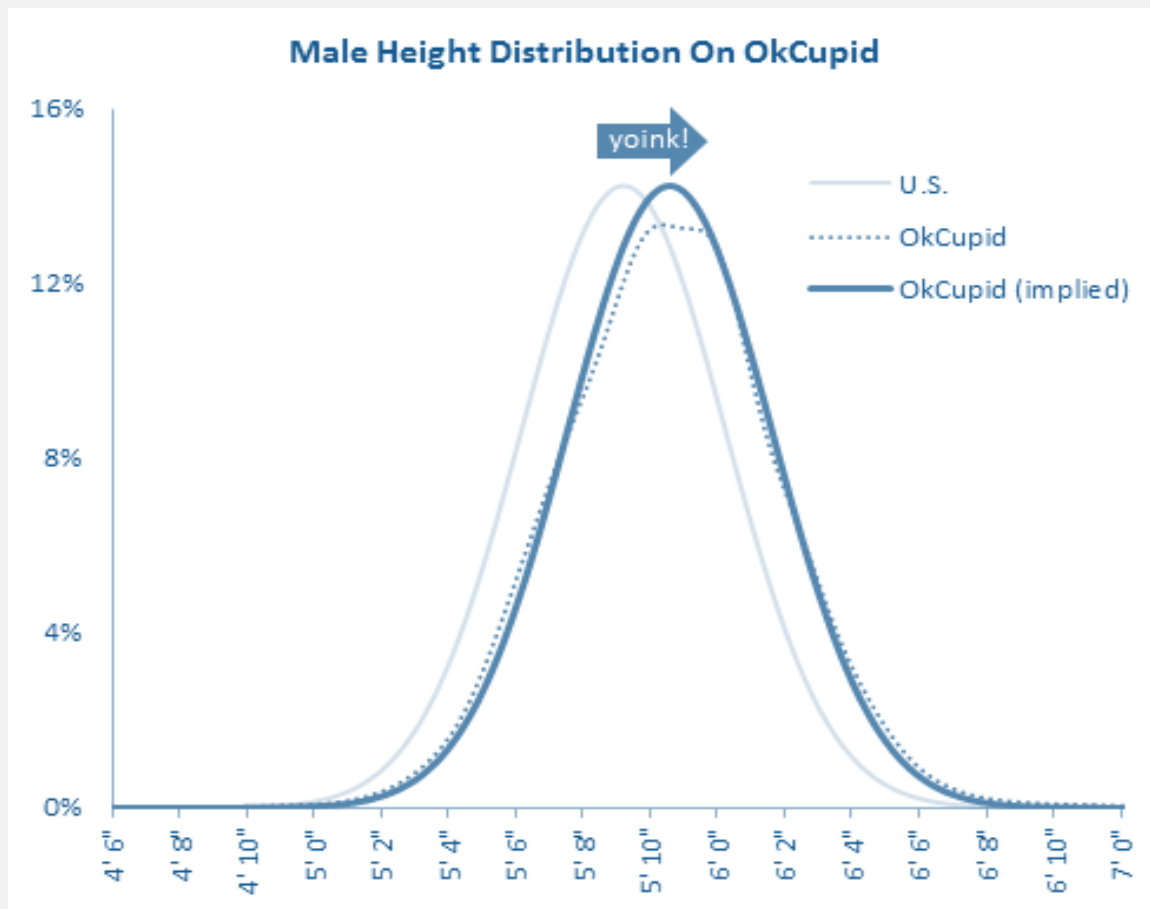


як перевірити чи розподіл є нормальним?



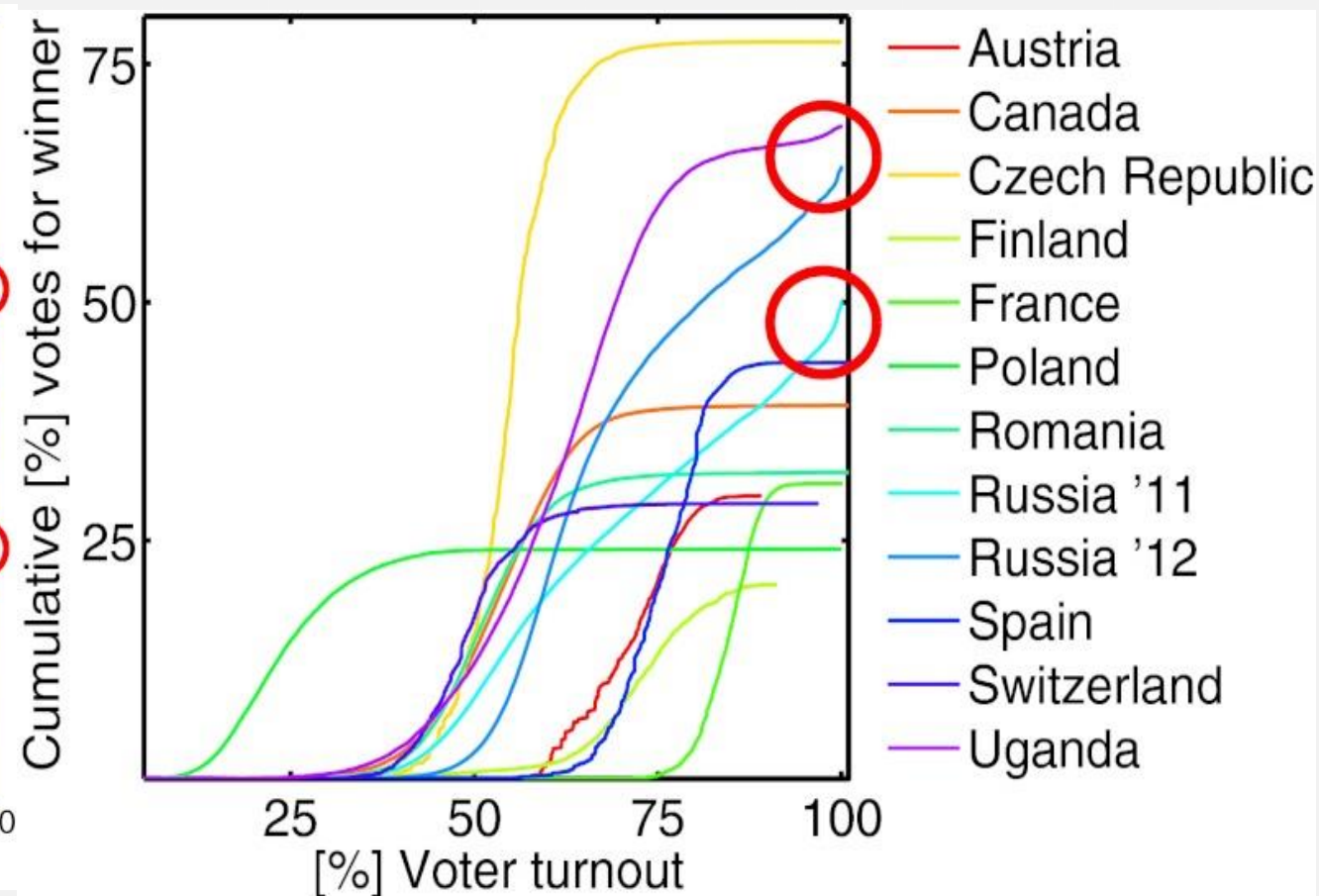
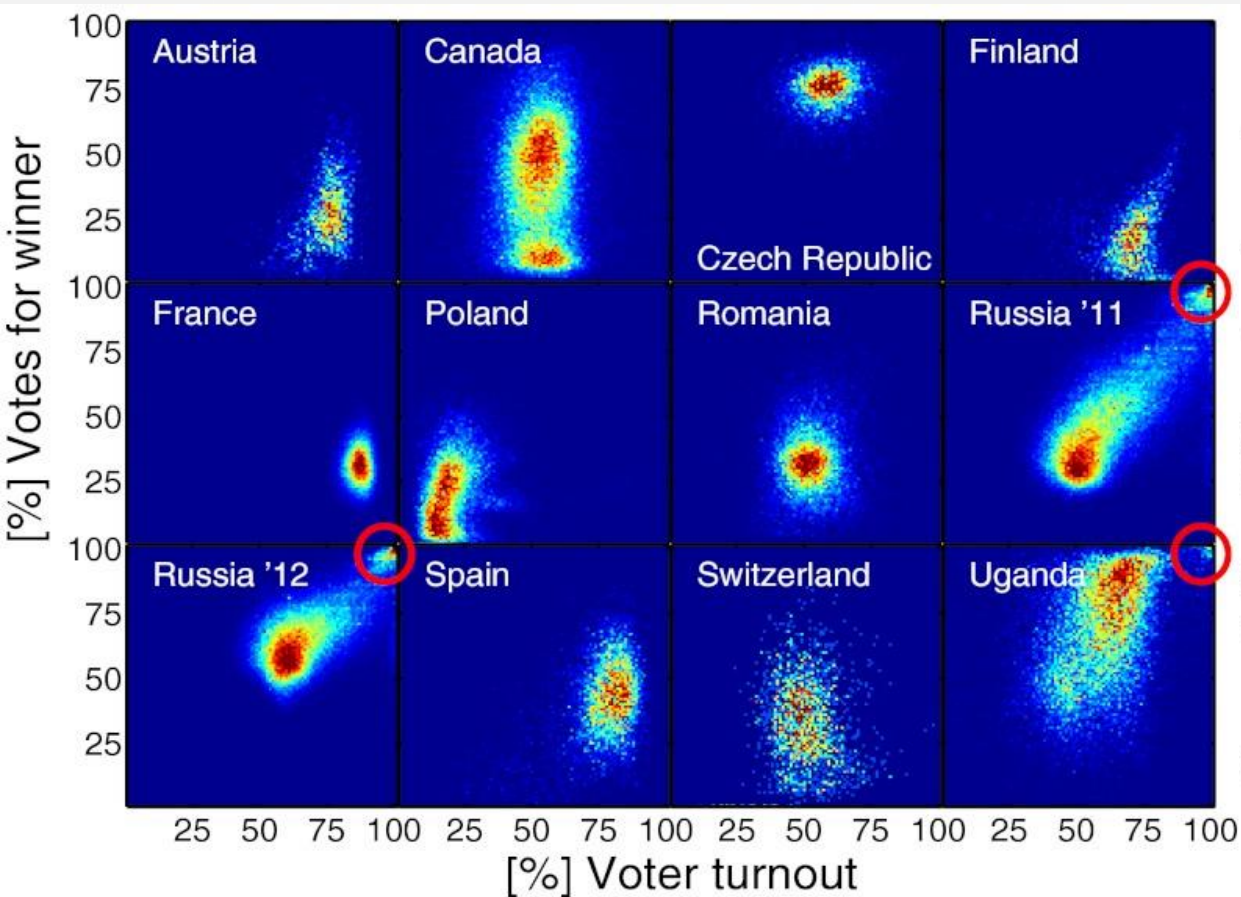
<http://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot>

викриття шахрайства



<https://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/>

Викриття шахрайства



<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3478593/>

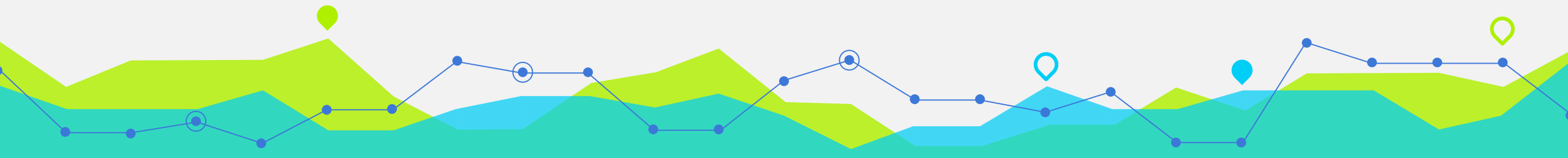


коваріація та
кореляція

коваріація і кореляція

Коваріація – міра лінійної залежності двох випадкових величин одна від одної.

Кореляція – зважена версія коваріації.

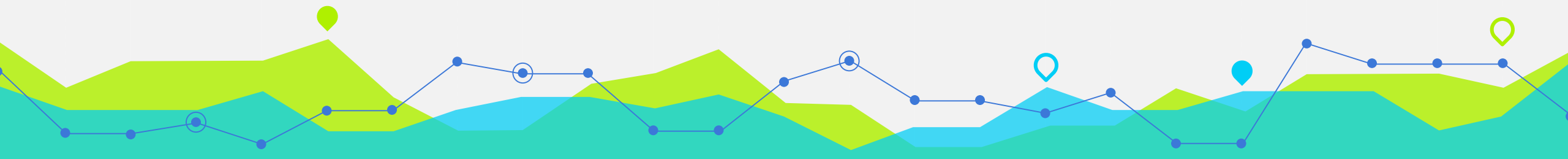


коефіцієнт Пірсона

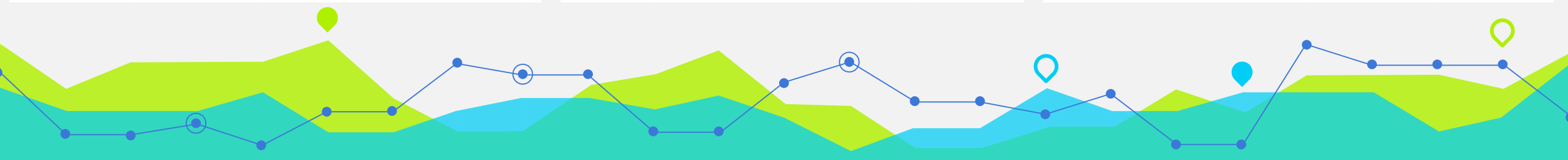
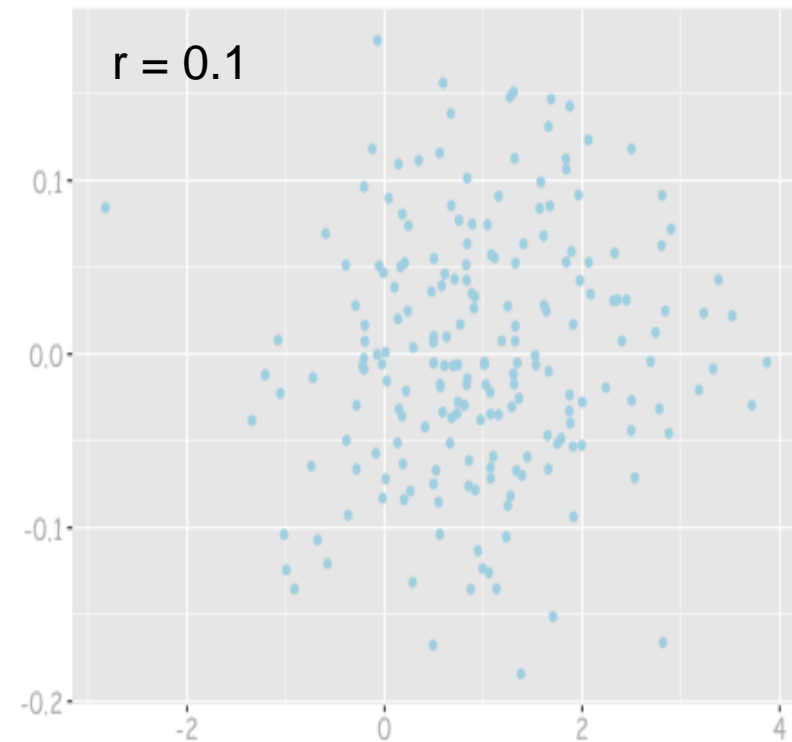
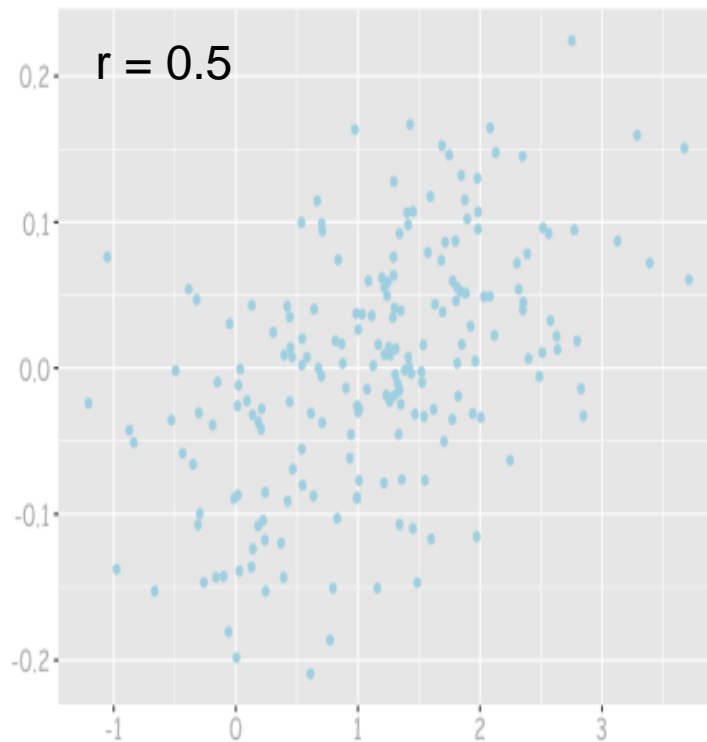
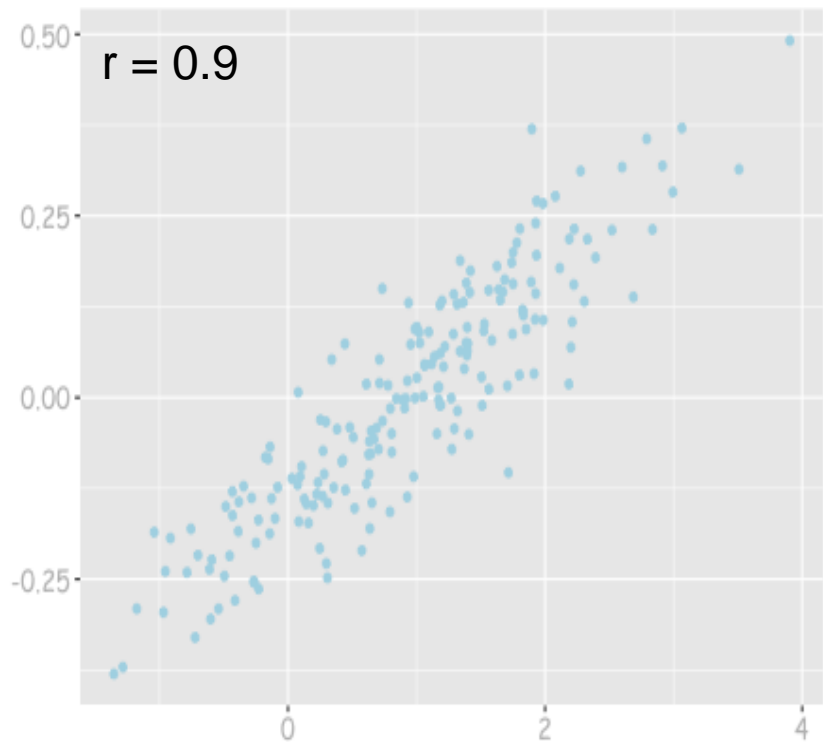
$$r = \frac{cov(x,y)}{S_x * S_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{\sum x_i}{N}$$

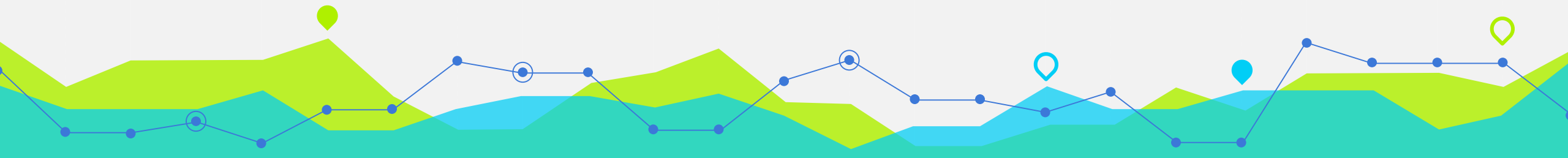
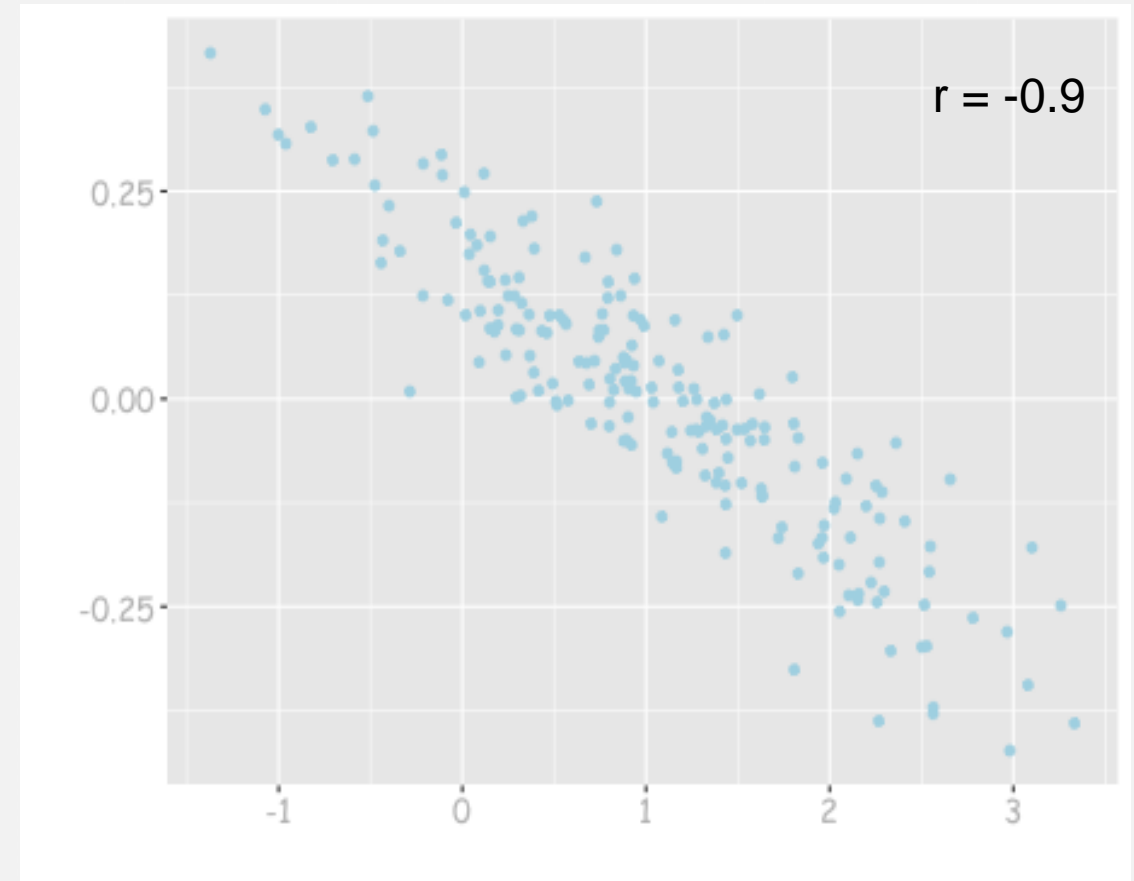
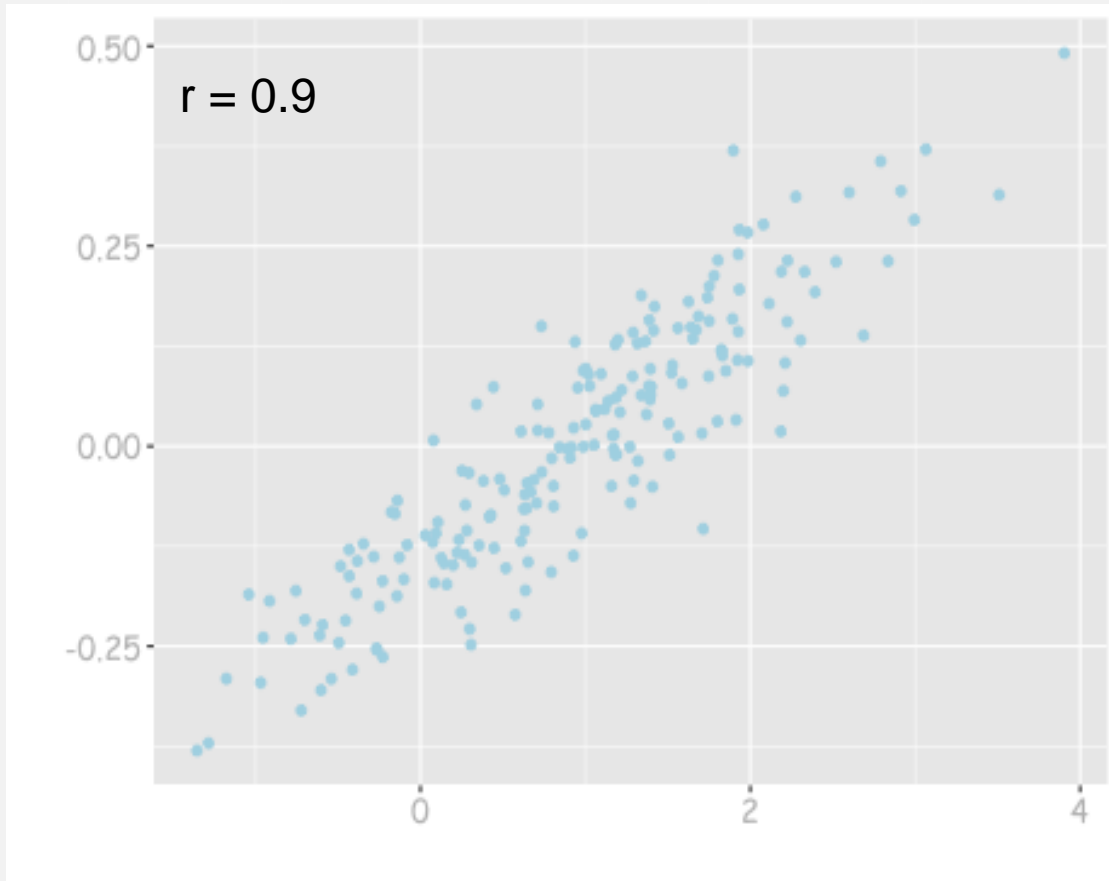
$$\bar{y} = \frac{\sum y_i}{N}$$



Абсолютне значення коефіцієнта кореляції дає уявлення про силу лінійного зв'язку між двома змінними

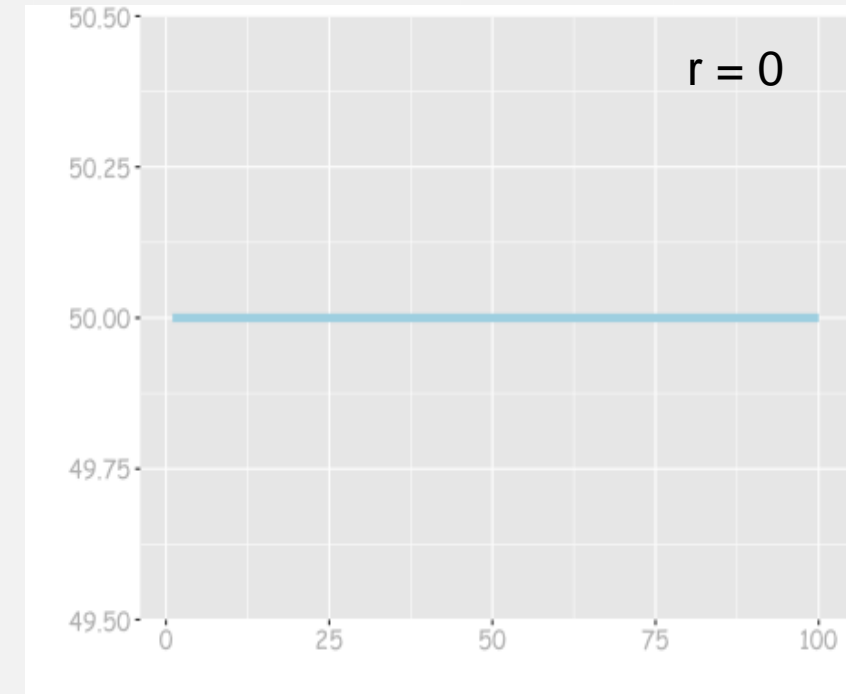
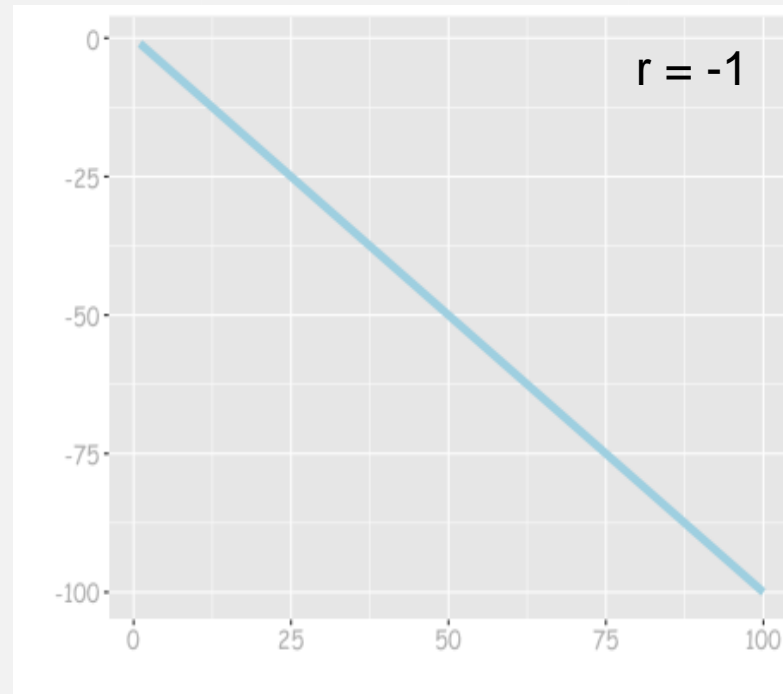
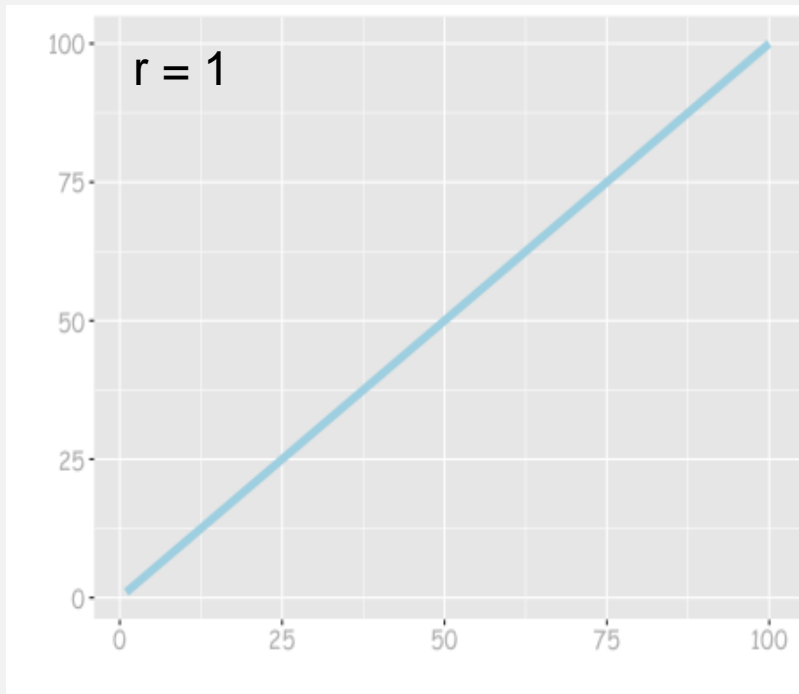


Знак коефіцієнта вказує напрямок зв'язку

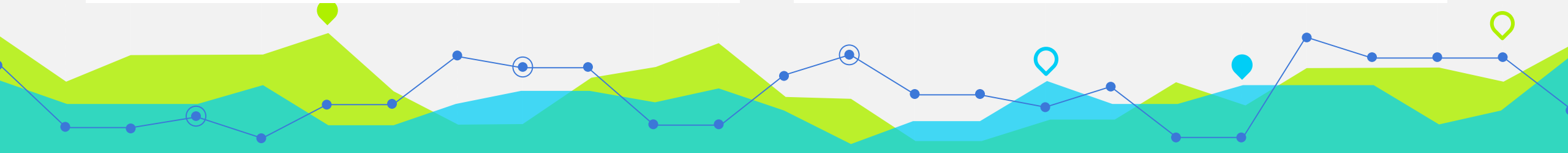
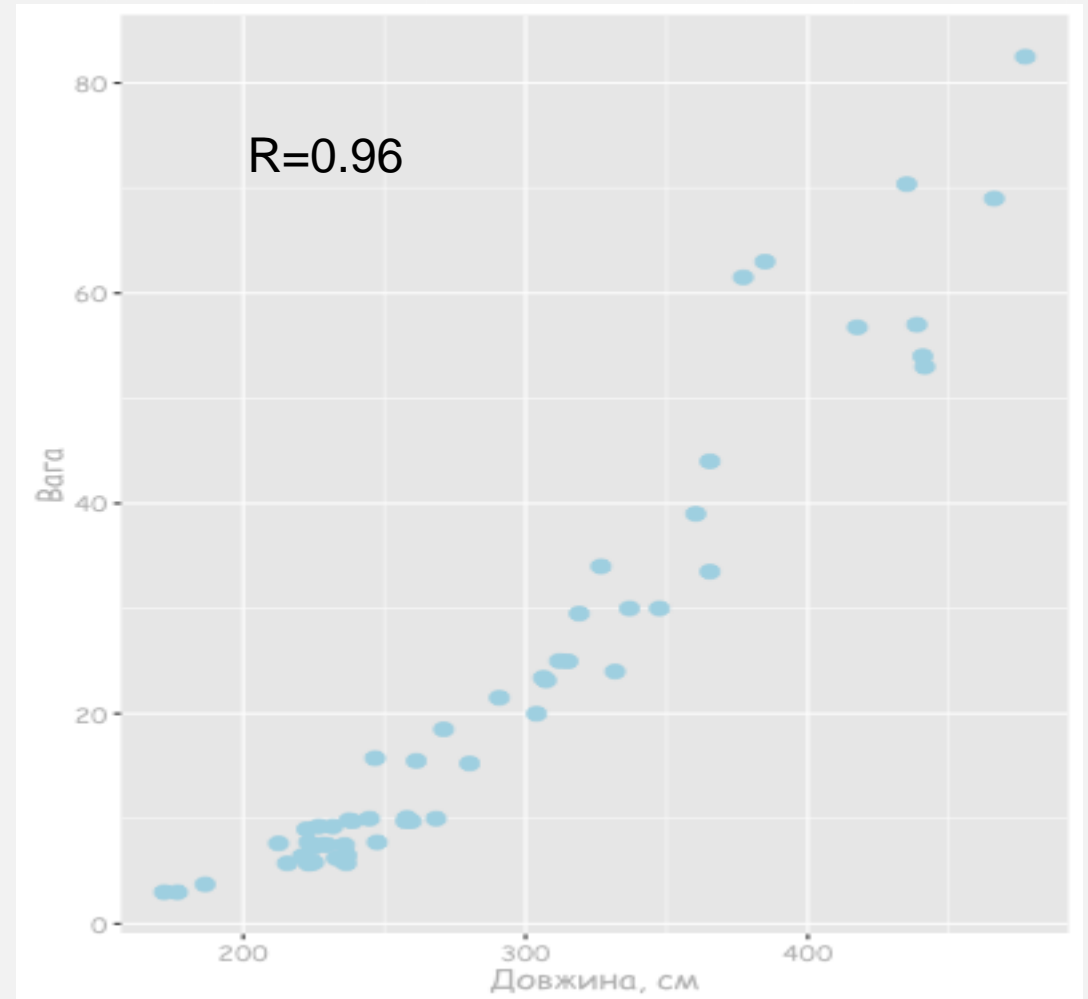
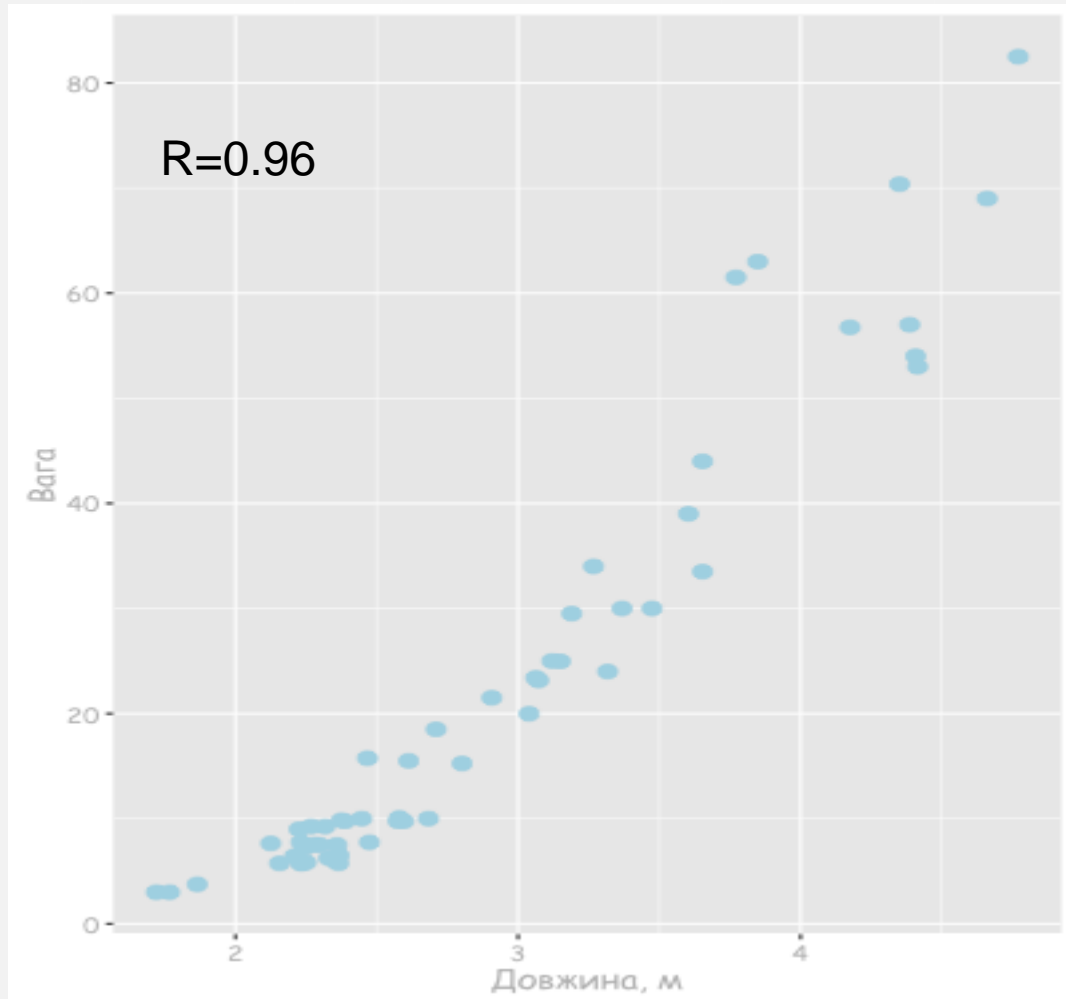


Коефіцієнт кореляції набуває значень між -1 та 1 вказує на сильну лінійну кореляцію.

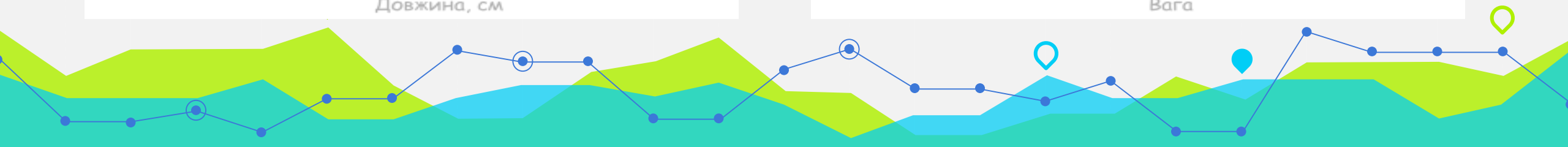
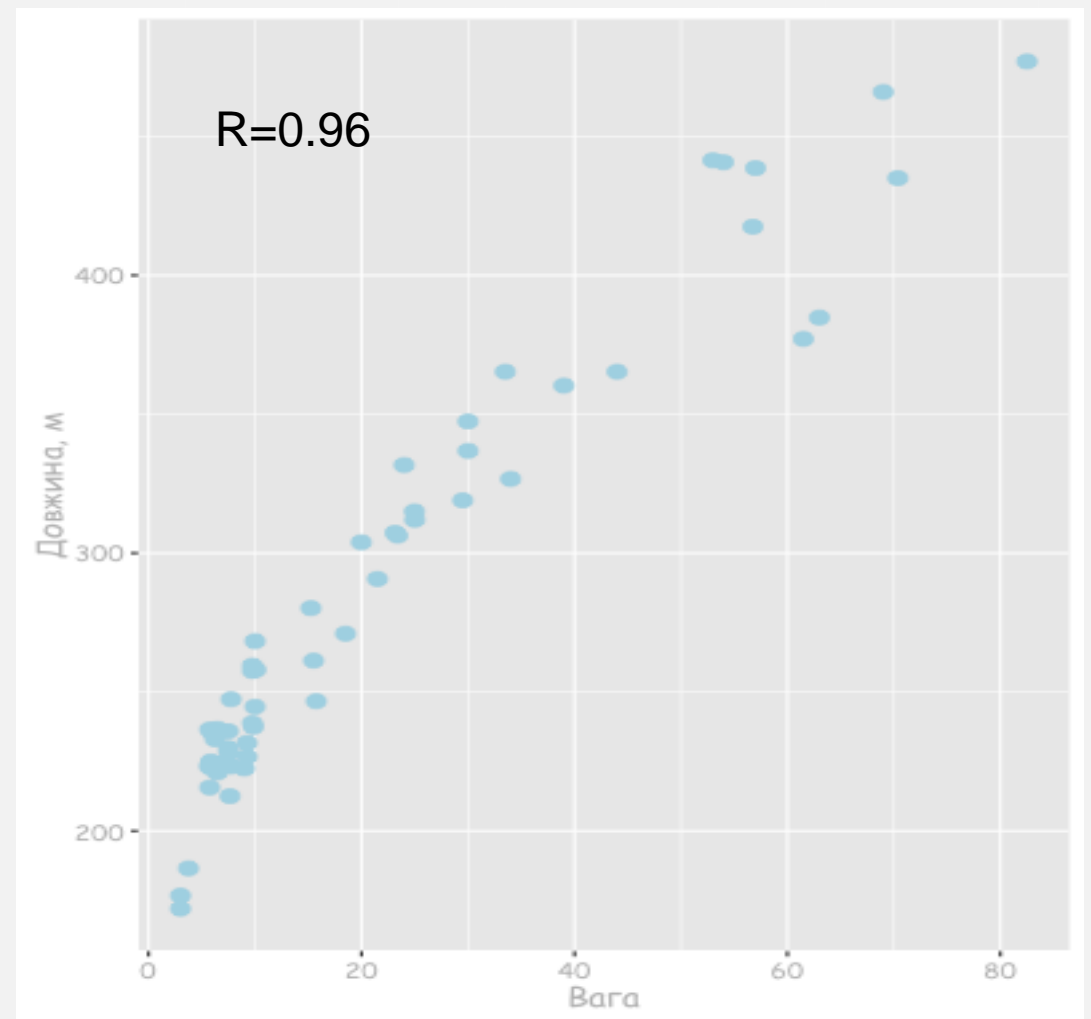
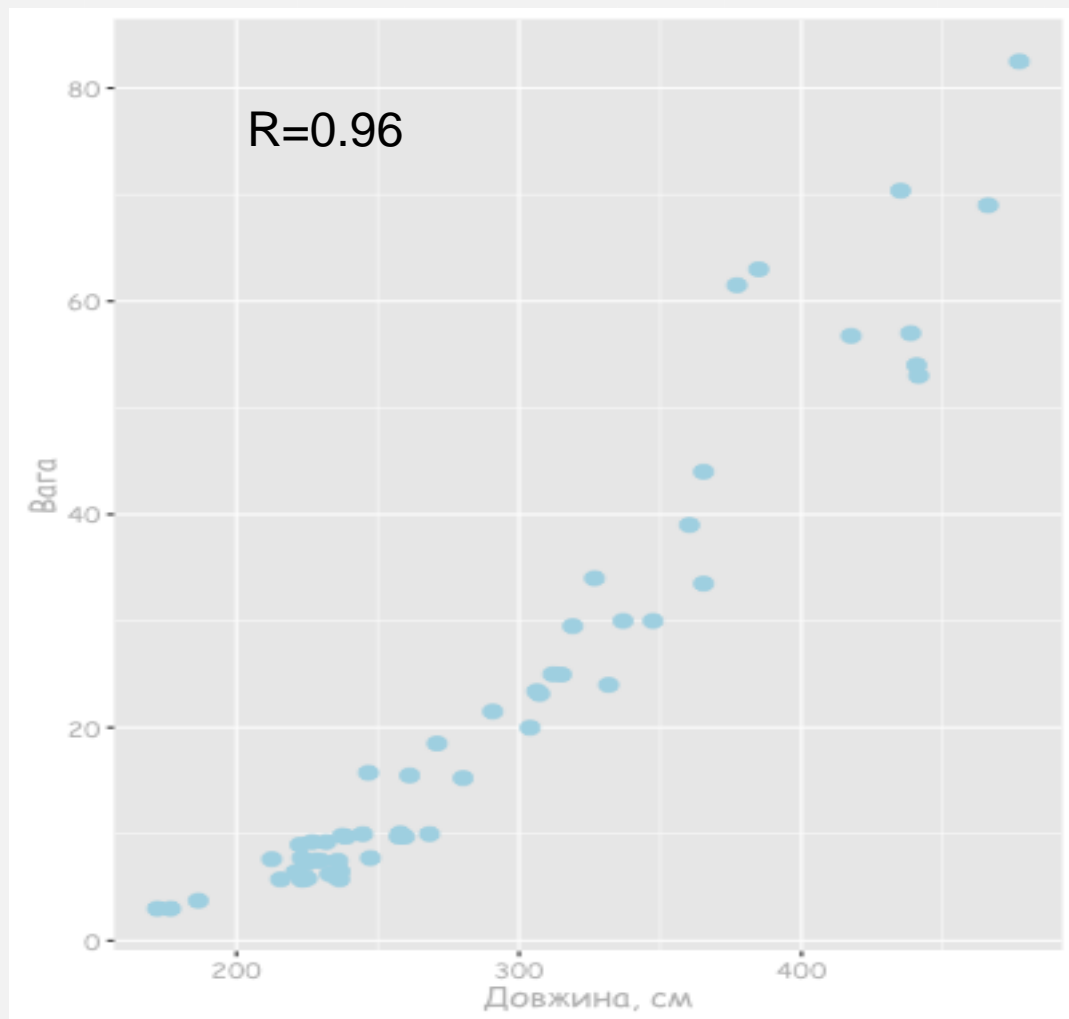
Значення коефіцієнта 0 вказує на відсутність лінійної кореляції



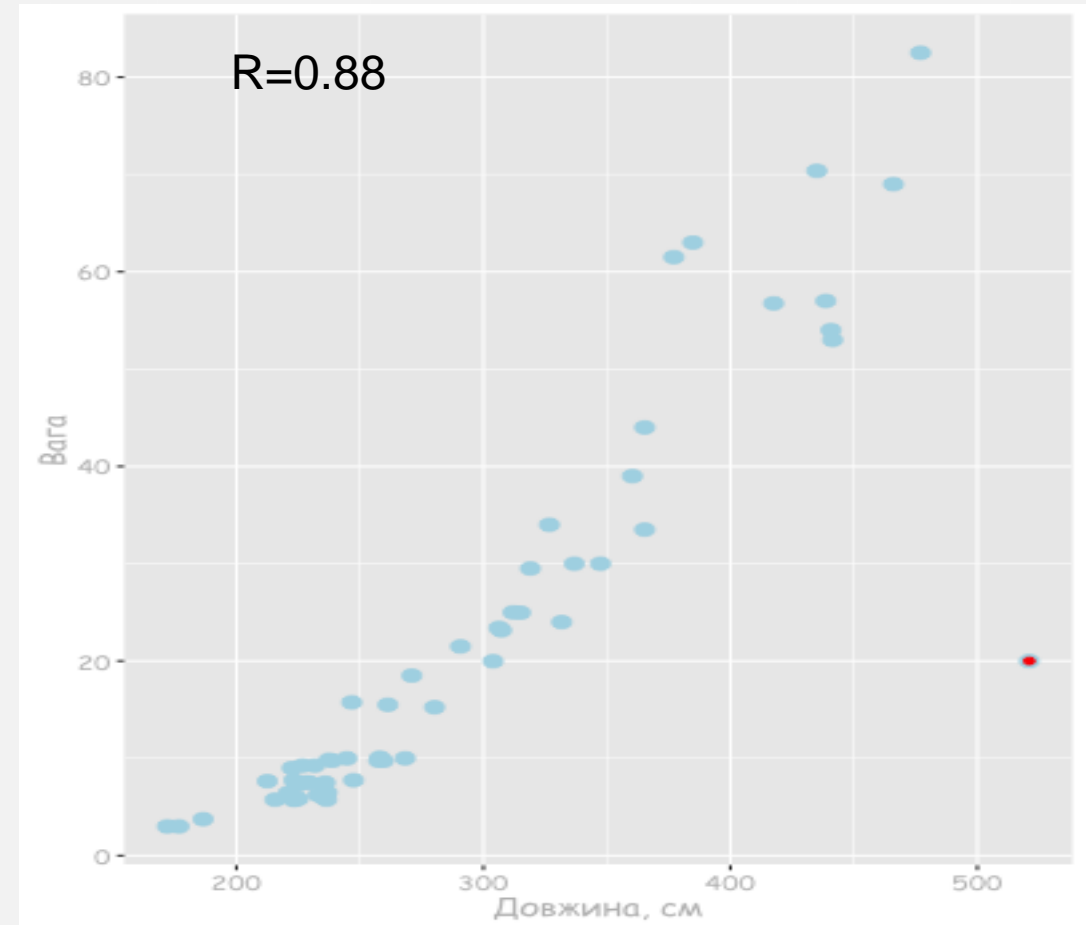
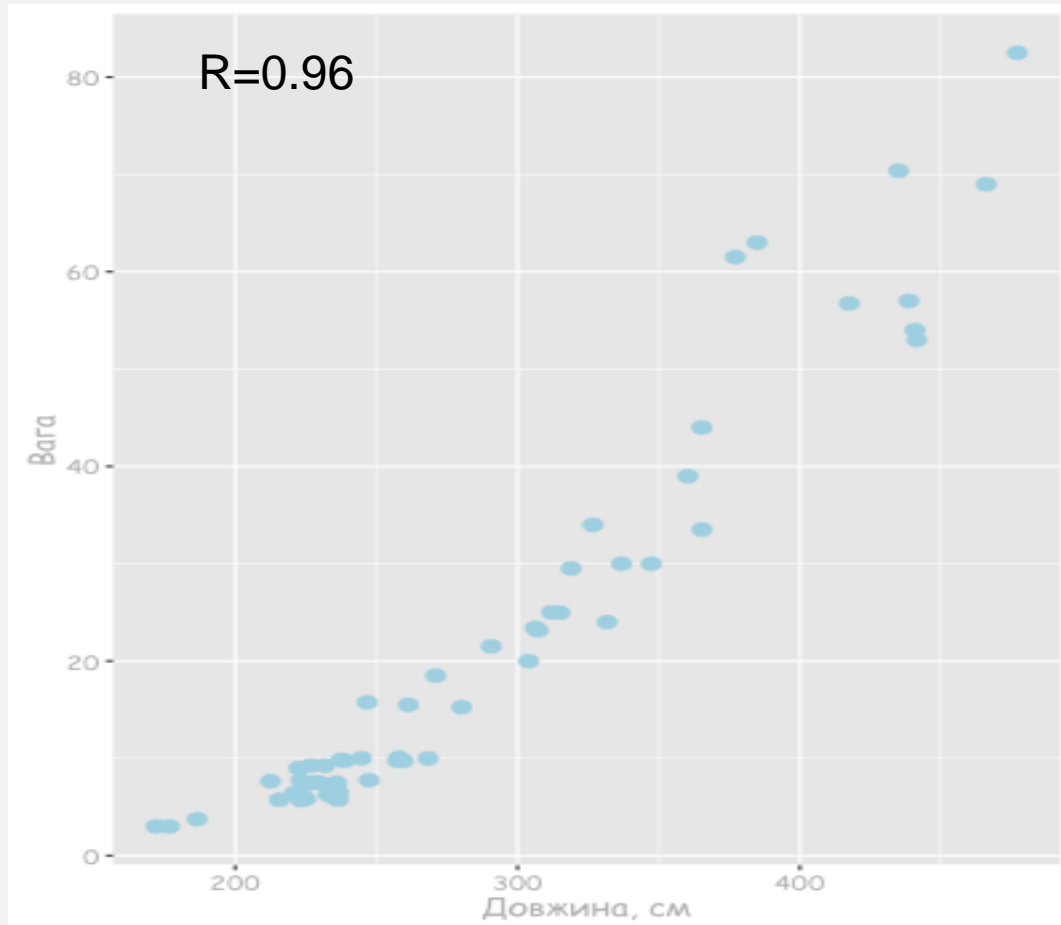
коефіцієнт кореляції не змінюється при зміні одиниць виміру



коефіцієнт кореляції є симетричним $r(x, y) = r(y, x)$



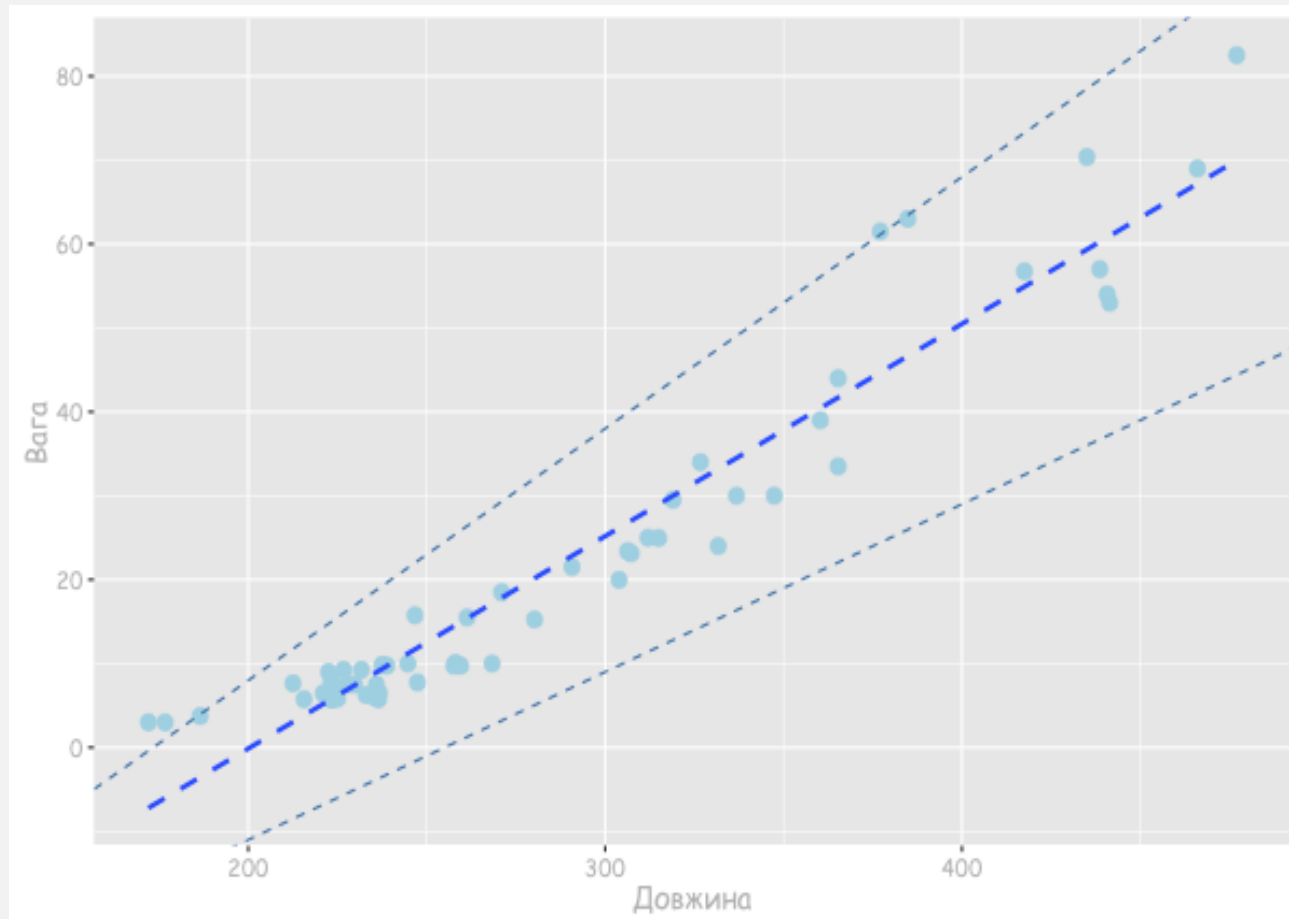
коефіцієнт кореляції чутливий до викидів





лінійна регресія

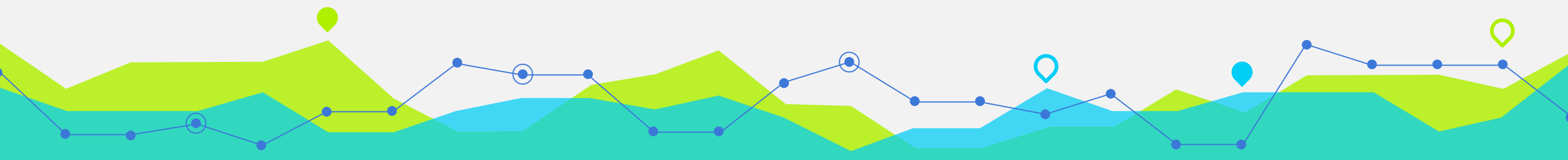
знаходження найкращої лінії

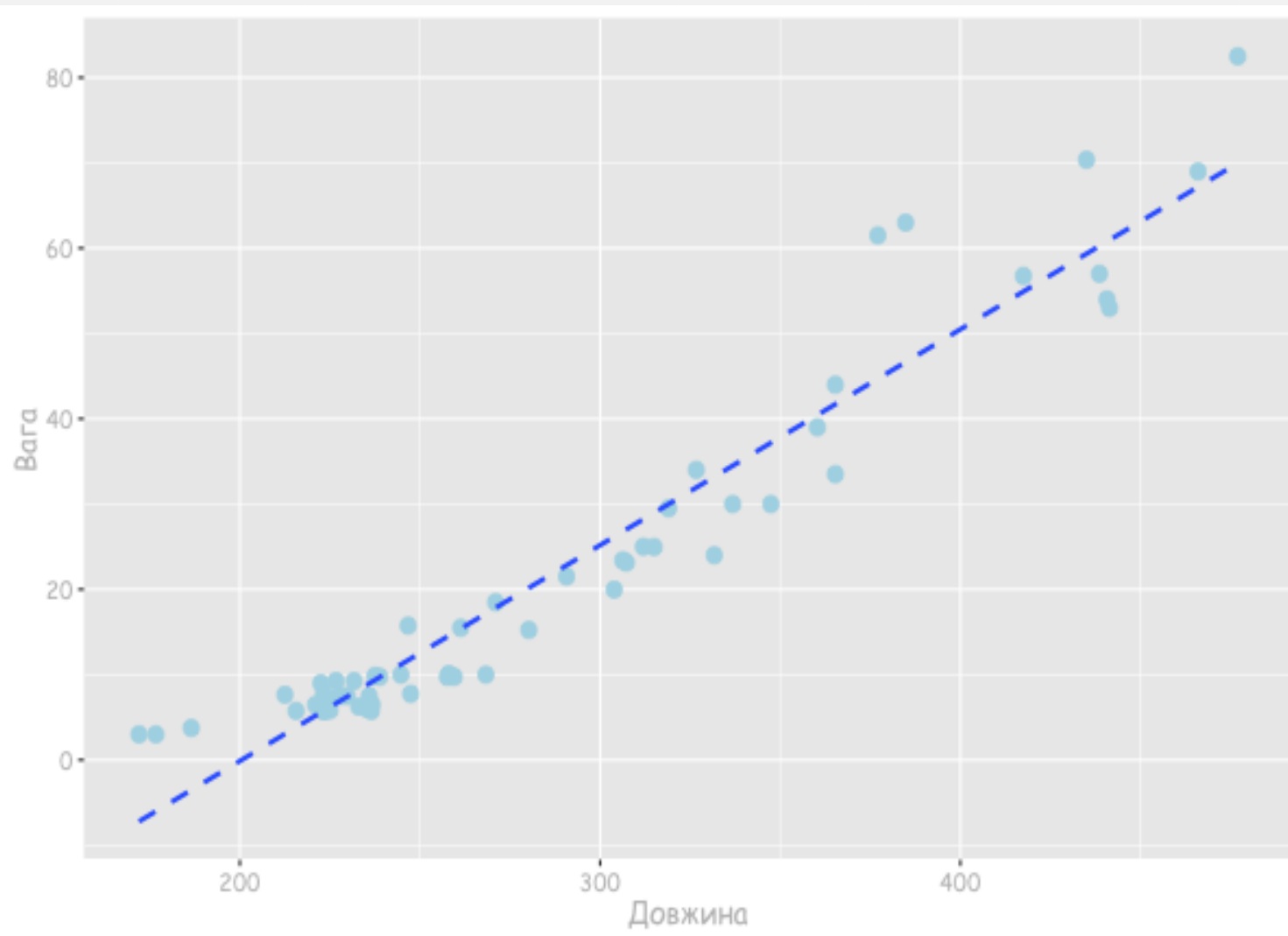


Мінімізуємо суму залишків
Мінімізуємо суму абсолютних
залишків
Мінімізуємо суму залишків у
квадраті.

$$\sum (y - \bar{y})^2$$

x – незалежна змінна
y – залежна змінна



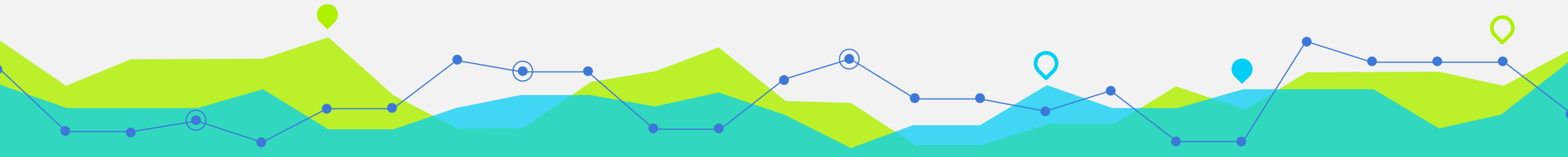


`cor(anaconda$V1, anaconda$V2) =`
0.96

$$y = 0.253 * x - 50.73$$

$$\hat{y} = ax + b$$

$$a = \frac{\sum (x - x_i)(y - y_i)}{\sum (x_i - \bar{x})^2} = r \frac{S_y}{S_x}$$



інтерпретація коефіцієнтів

Call:

```
lm(formula = anaconda$V2 ~ anaconda$V1)
```

Residuals:

Min	1Q	Median	3Q	Max
-9.2050	-3.9127	-0.2454	1.9430	16.8067

Coefficients:

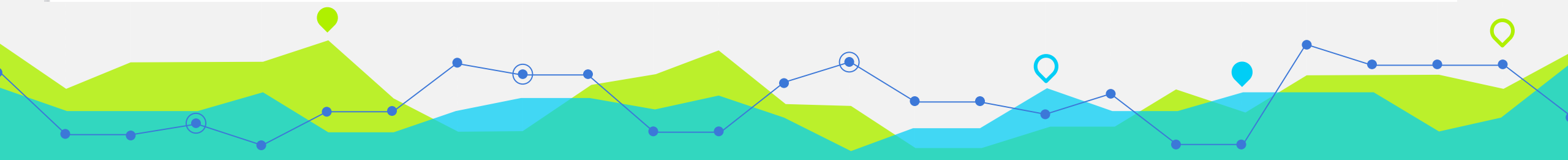
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-50.730584	2.946034	-17.22	<2e-16	***
anaconda\$V1	0.253047	0.009857	25.67	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.754 on 54 degrees of freedom

Multiple R-squared: 0.9243, Adjusted R-squared: 0.9229

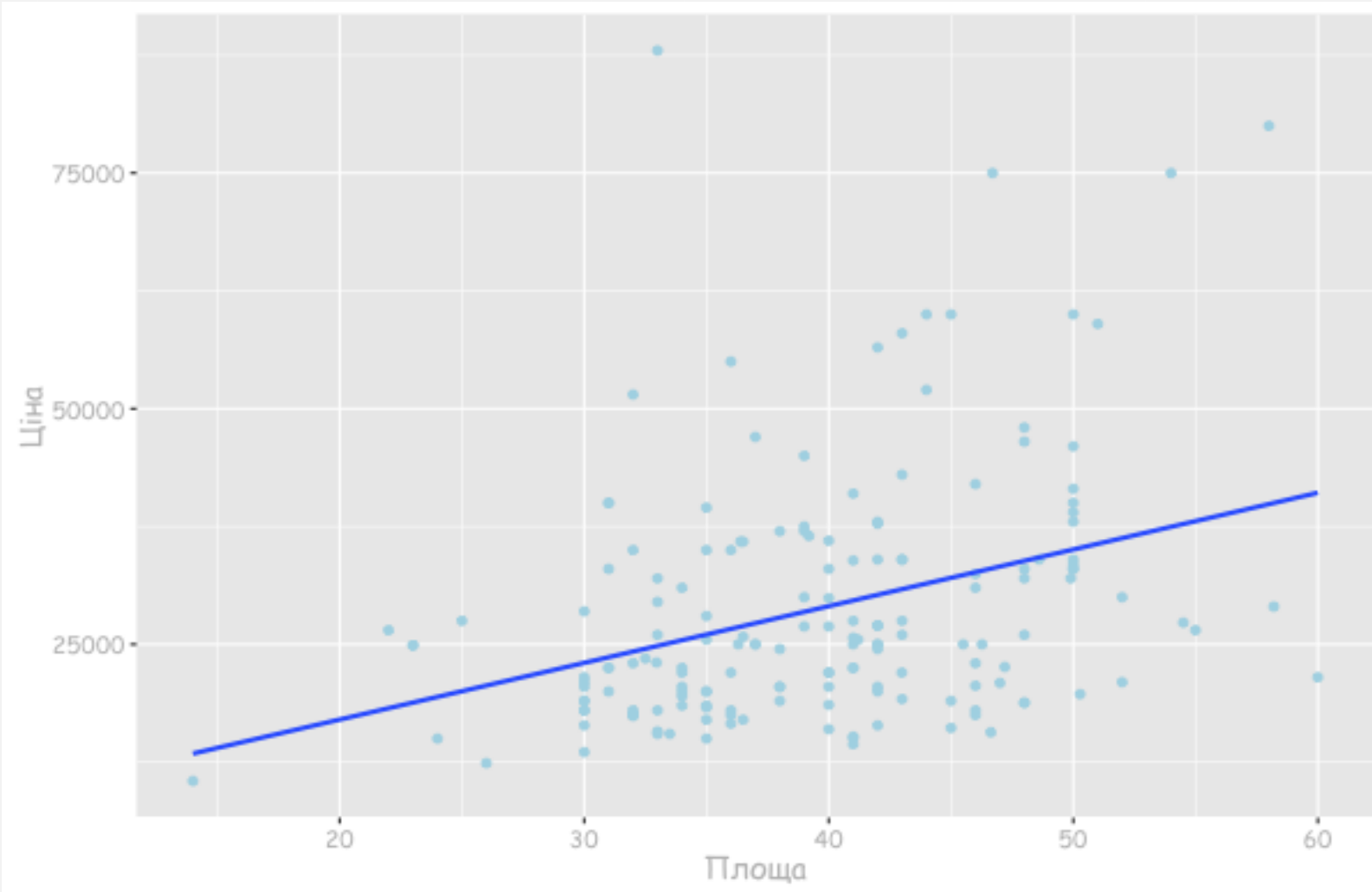
F-statistic: 659 on 1 and 54 DF, p-value: < 2.2e-16



R^2

- Використовується як оцінка сили лінійної залежності між незалежною та залежною змінними
- Має значення від 0 до 1
- Рахується як квадрат коефіцієнта кореляції
- Говорить, який відсоток варіативності залежної змінної пояснюється лінійною моделлю
- Залишок пояснюється змінними, які не включені в модель

квартири



$$\text{Ціна} = 4988.3 + 601.5 * \text{площа}$$

Call:

```
lm(formula = one_room_flats$Ціна_usd ~ one_room_flats$Загальна_площа)
```

Residuals:

Min	1Q	Median	3Q	Max
-19576	-7690	-2970	5949	63163

Coefficients:

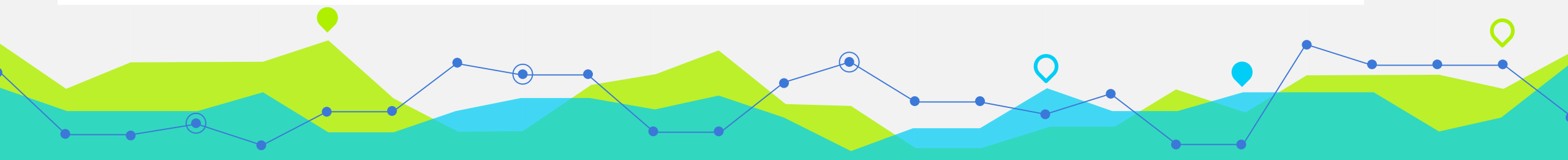
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4988.3	5034.3	0.991	0.323
one_room_flats\$Загальна_площа	601.5	125.1	4.806	0.0000034 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12510 on 168 degrees of freedom

Multiple R-squared: 0.1209, Adjusted R-squared: 0.1156

F-statistic: 23.1 on 1 and 168 DF, p-value: 0.000003396



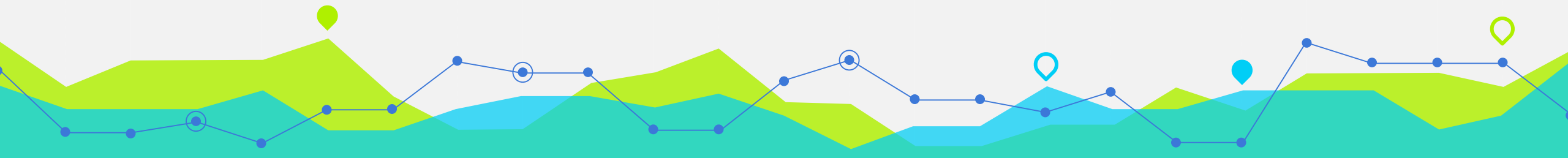
УМОВИ

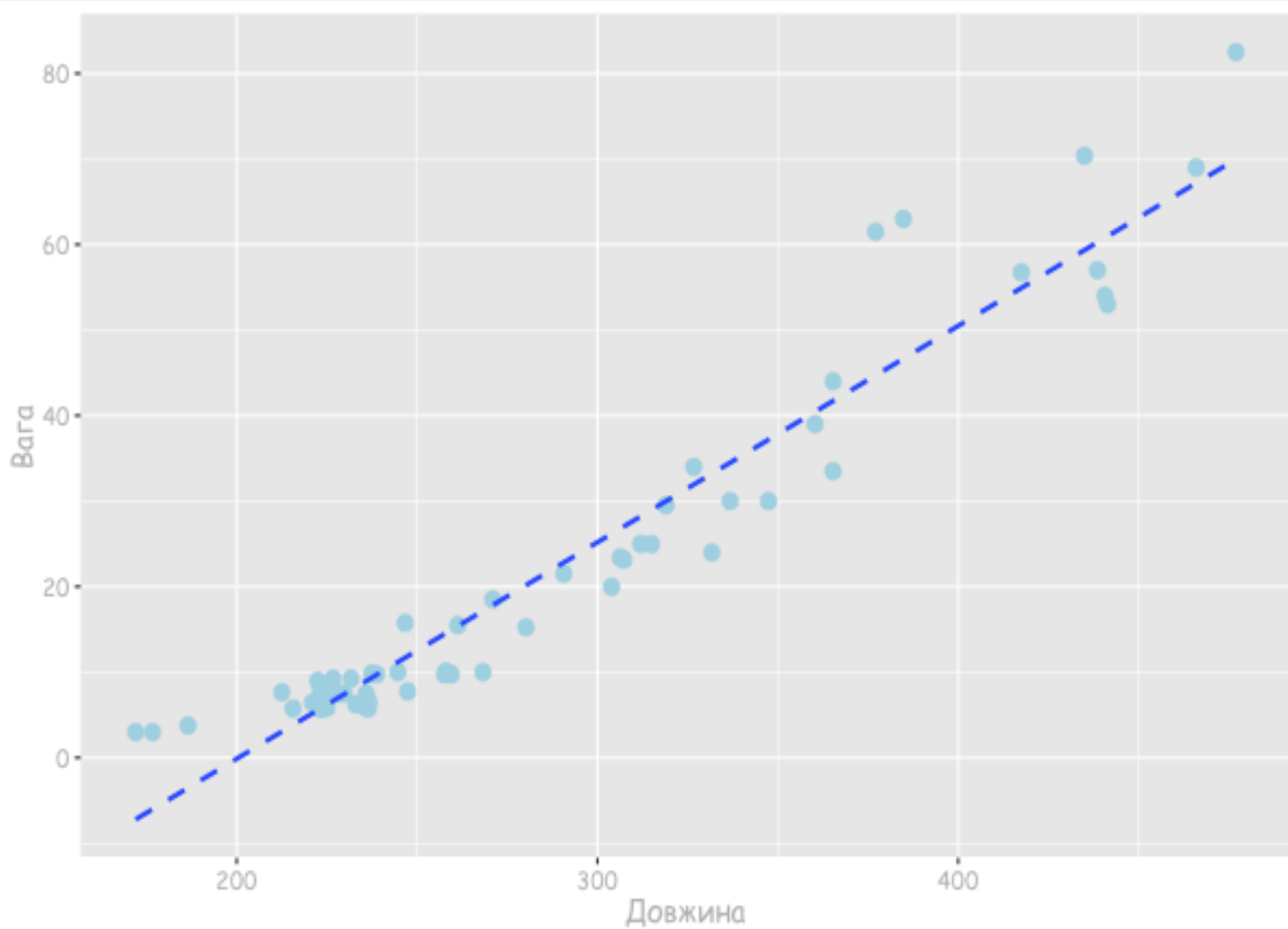
Лінійність

Нормальний
розподіл залишків

Гомоскедастичність (стала
варіативність
залишків)

https://gallery.shinyapps.io/slr_diag/





залишком спостережуваної змінної є різниця між цим спостережуваним значенням та оцінкою значення досліджуваної величини

Залишок:

$$e_i = y_i - \hat{y}_i$$

Недооцінка
Переоцінка



Diagnostics for simple linear regression

Select a trend:

- ☒ Linear up
- ☐ Linear down
- ☐ Curved up
- ☐ Curved down
- ☐ Fan-shaped

☐ Show residuals

This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

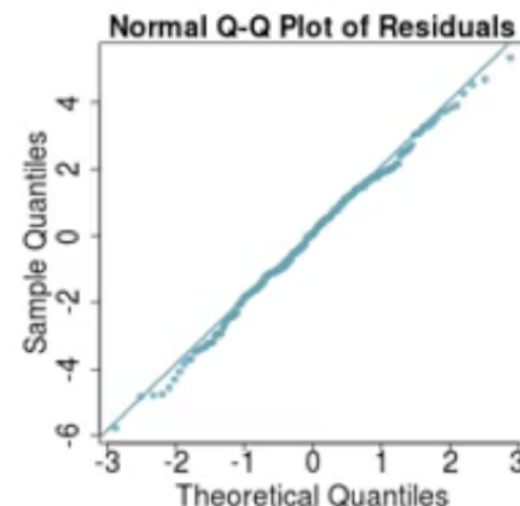
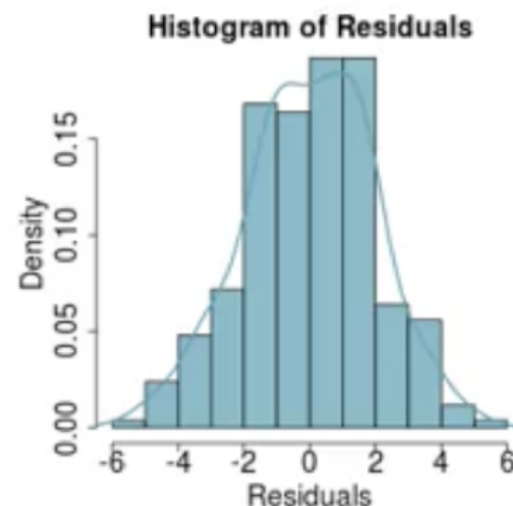
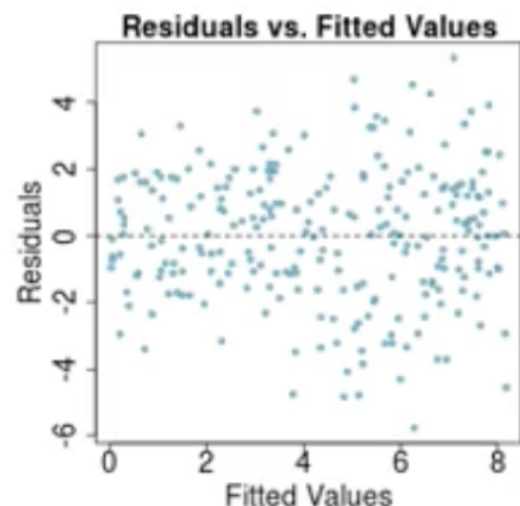
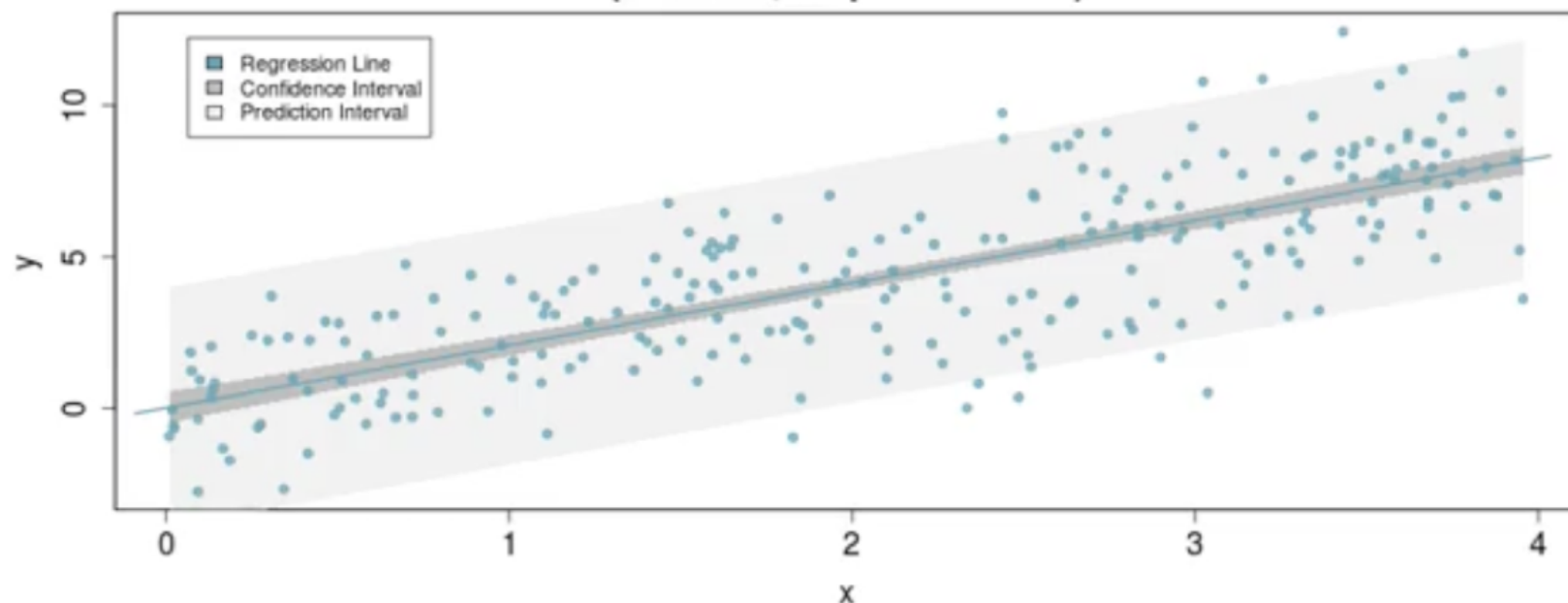
[Rate this app!](#)

[View code](#)

[Check out other apps](#)

[Want to learn more for free?](#)

Regression Model
($R = 0.7718$, $R\text{-squared} = 0.5956$)



Diagnostics for simple linear regression

Select a trend:

- ☐ Linear up
- ☐ Linear down
- ☒ Curved up
- ☐ Curved down
- ☐ Fan-shaped

☐ Show residuals

This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

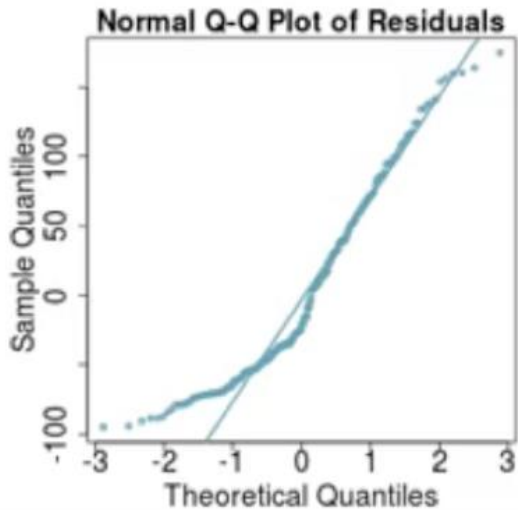
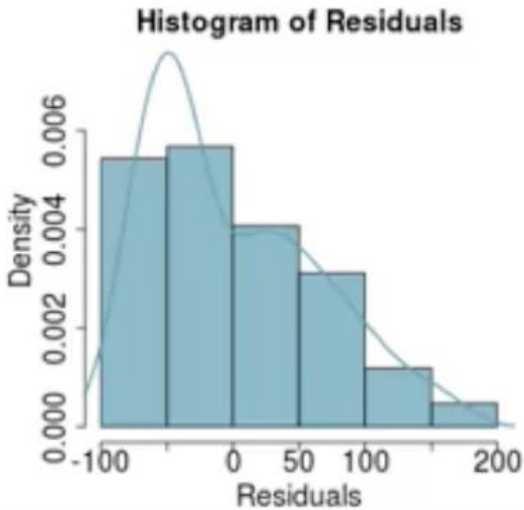
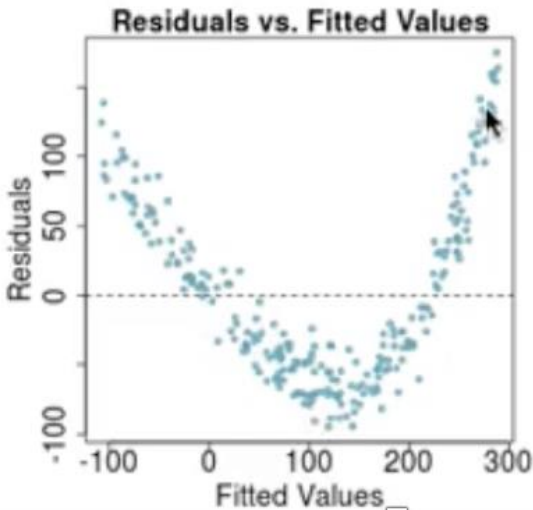
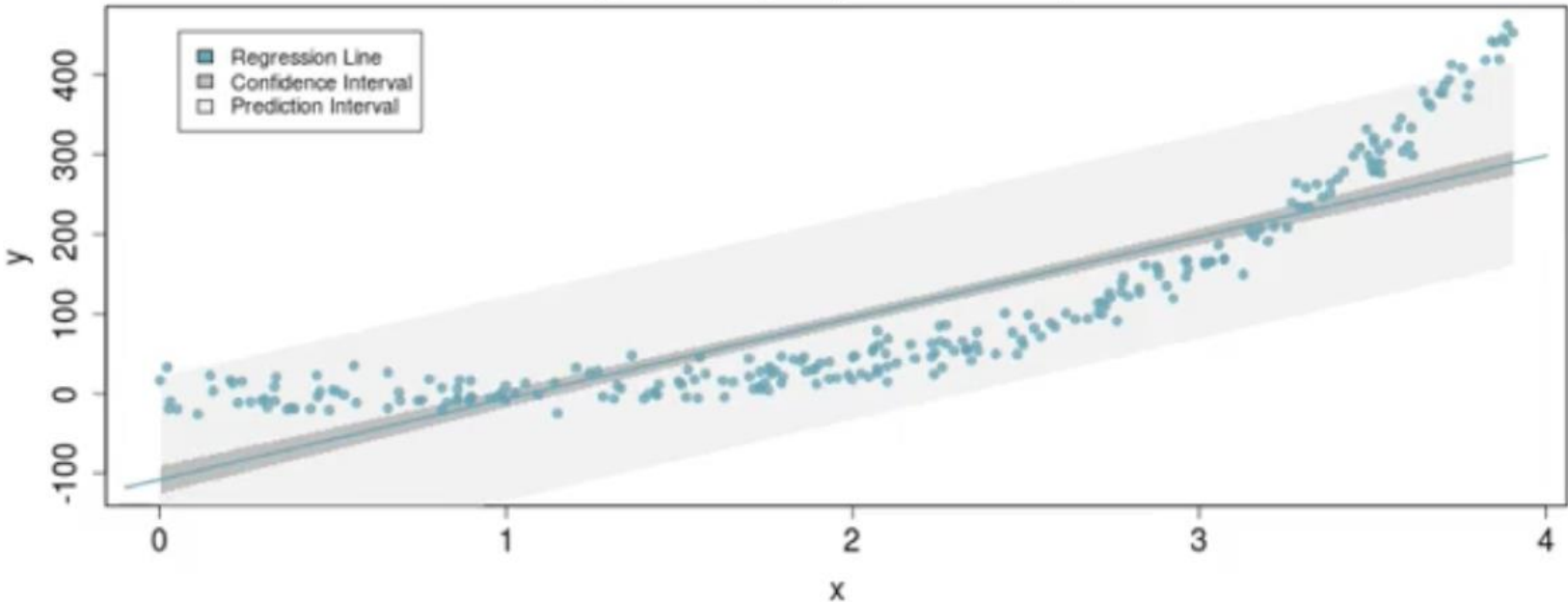
[Rate this app!](#)

[View code](#)

[Check out other apps](#)

[Want to learn more for free?](#)

Regression Model
($R = 0.8663$, $R\text{-squared} = 0.7504$)



Diagnostics for simple linear regression

Select a trend:

- ☐ Linear up
- ☐ Linear down
- ☐ Curved up
- ☒ Curved down
- ☐ Fan-shaped

☐ Show residuals

This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

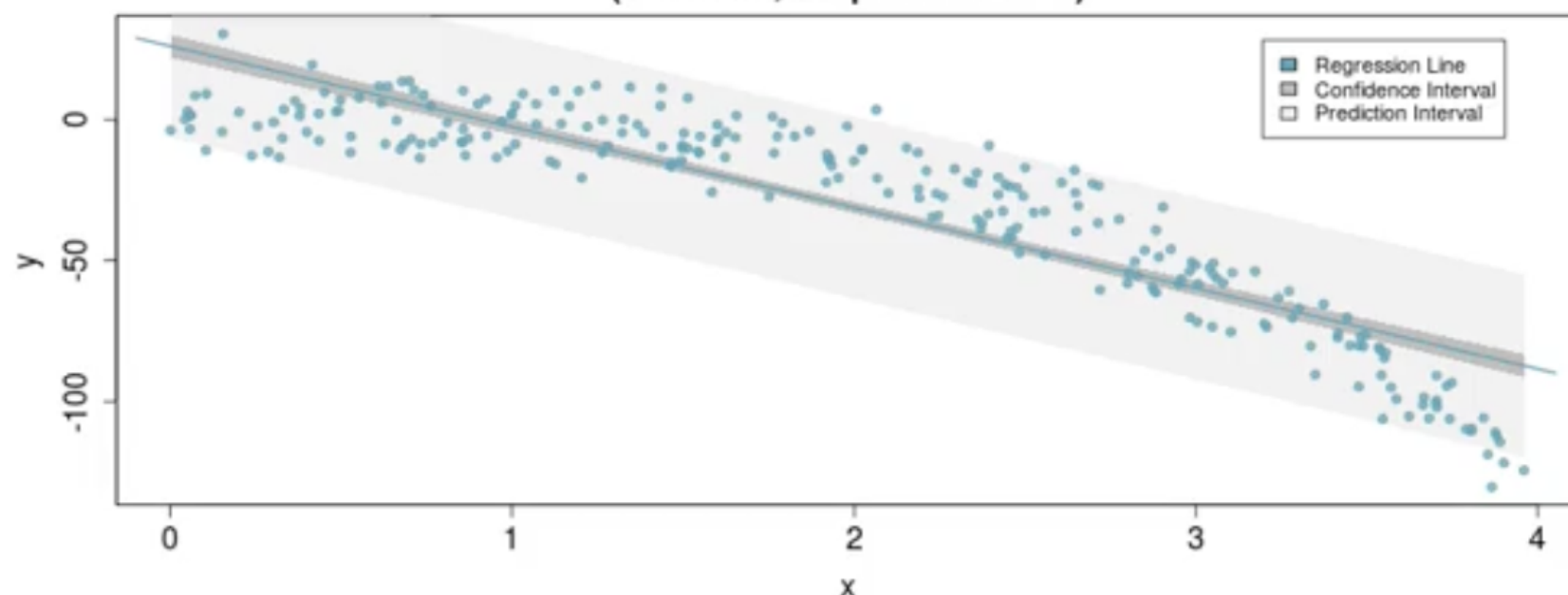
[Rate this app!](#)

[View code](#)

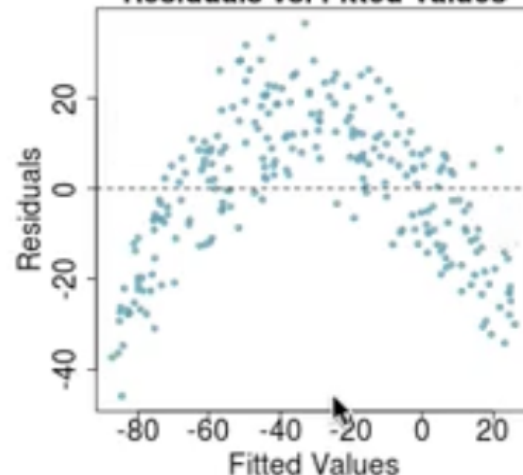
[Check out other apps](#)

[Want to learn more for free?](#)

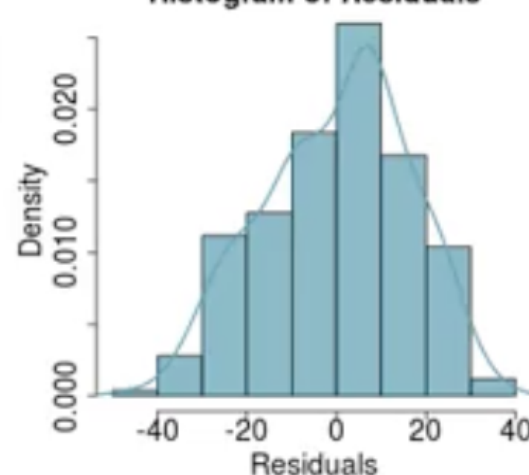
Regression Model
($R = 0.8949$, $R\text{-squared} = 0.8009$)



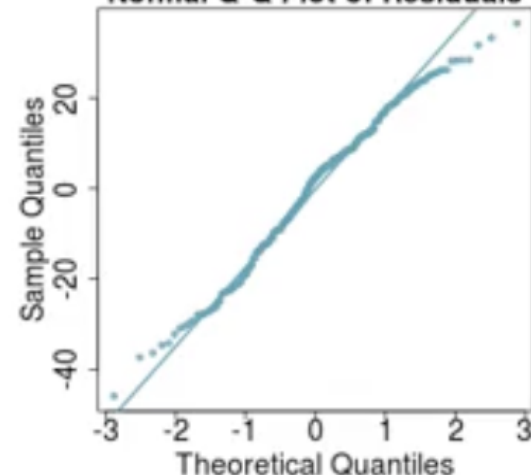
Residuals vs. Fitted Values



Histogram of Residuals



Normal Q-Q Plot of Residuals



Diagnostics for simple linear regression

Select a trend:

- ☐ Linear up
- ☐ Linear down
- ☐ Curved up
- ☐ Curved down
- ☒ Fan-shaped

☐ Show residuals

This applet uses ordinary least squares (OLS) to fit a regression line to the data with the selected trend. The applet is designed to help you practice evaluating whether or not the linear model is an appropriate fit to the data. The three diagnostic plots on the lower half of the page are provided to help you identify undesirable patterns in the residuals that may arise from non-linear trends in the data.

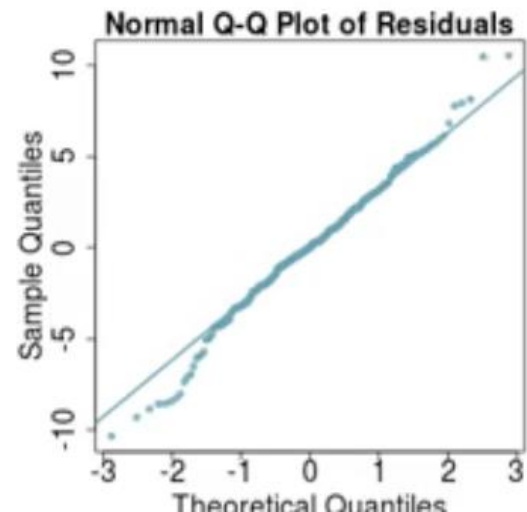
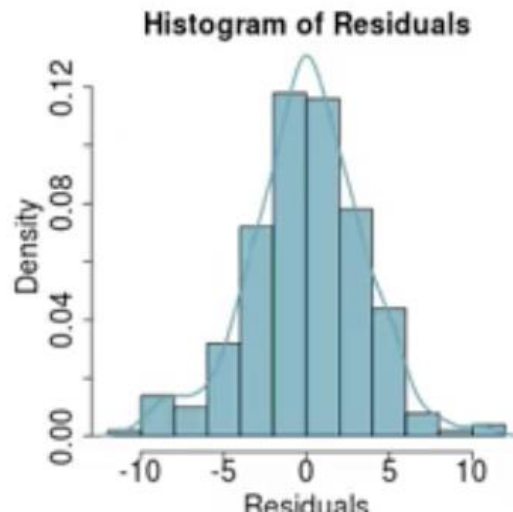
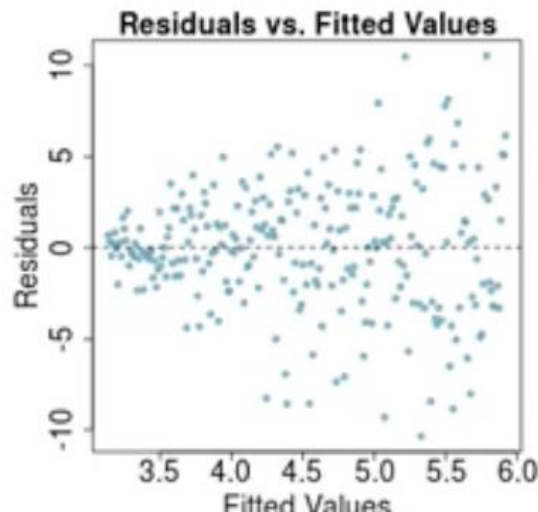
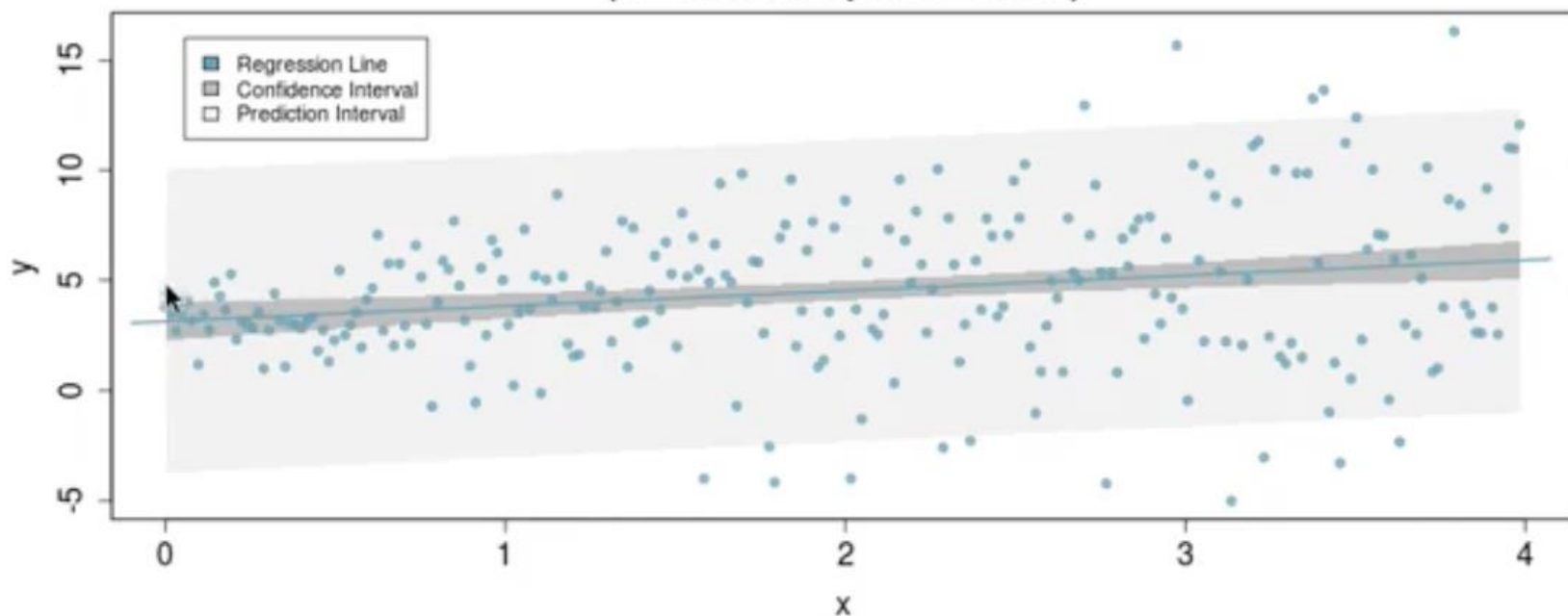
[Rate this app!](#)

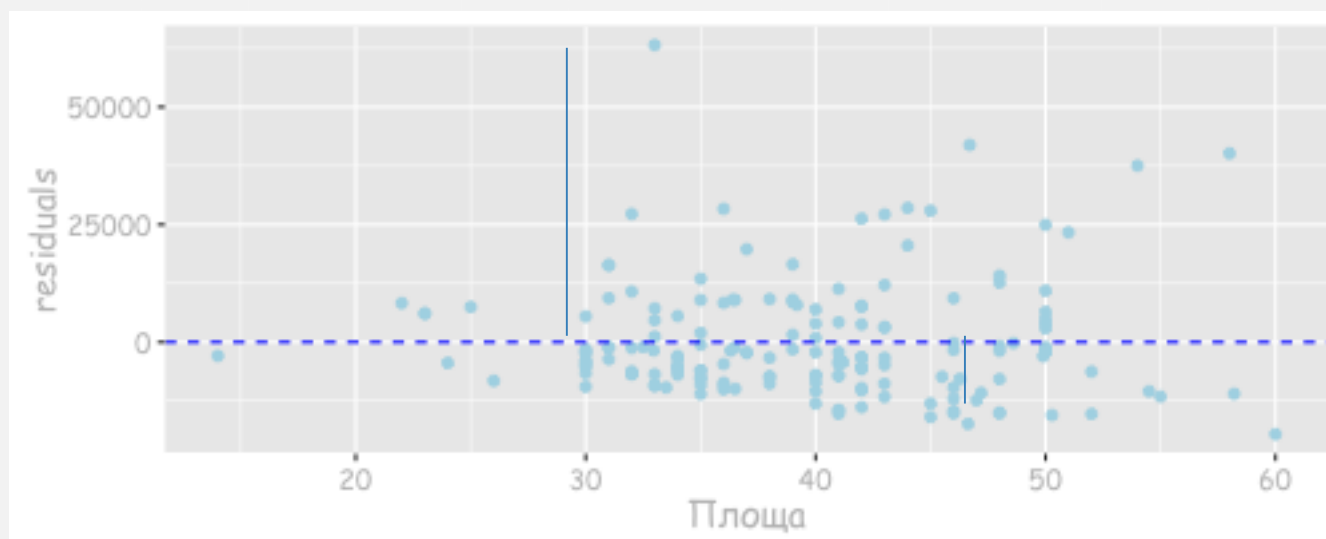
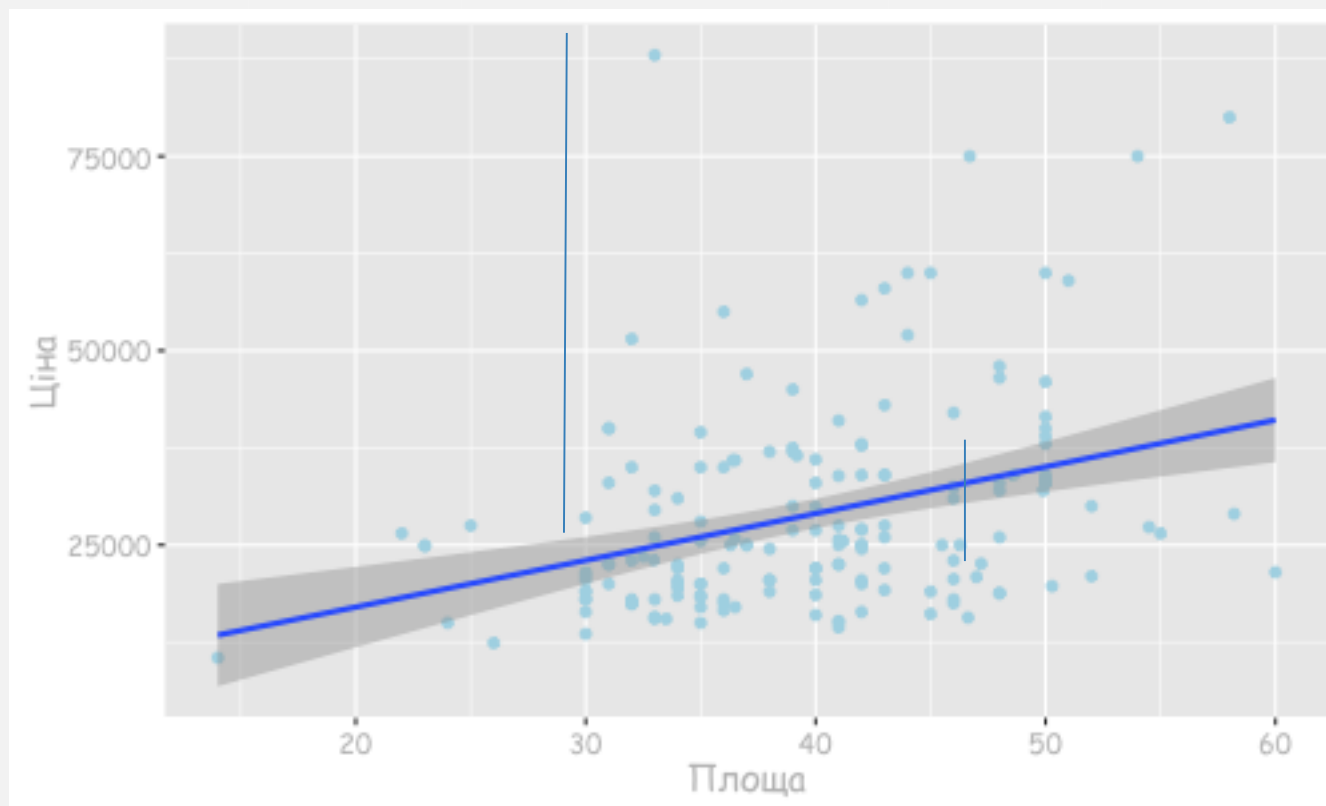
[View code](#)

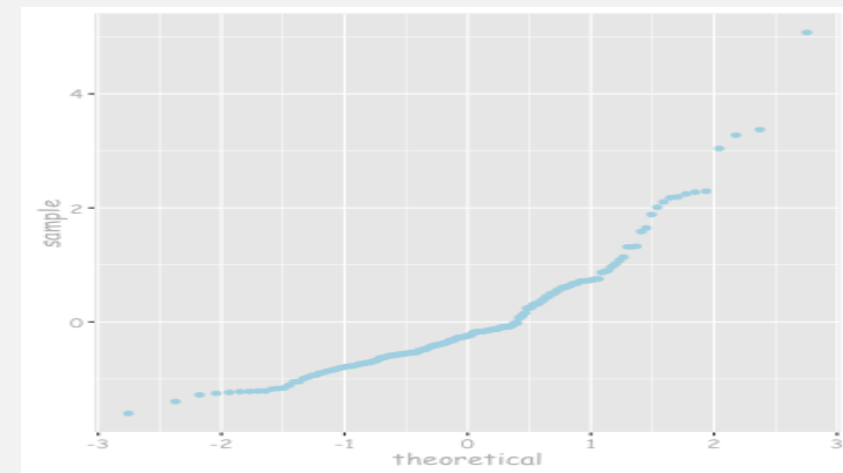
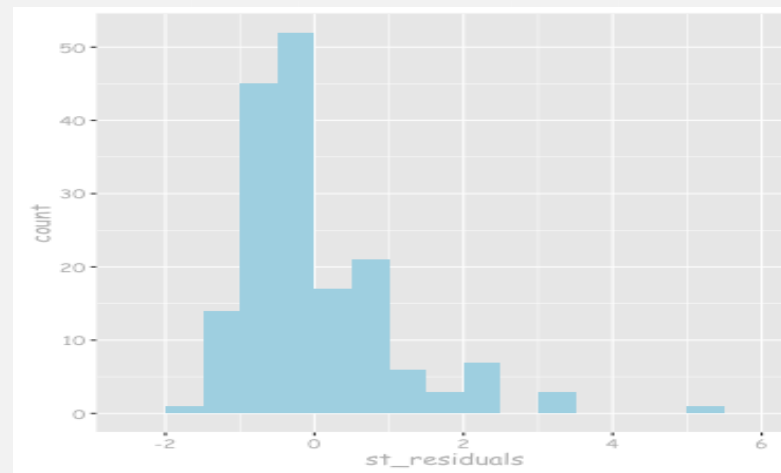
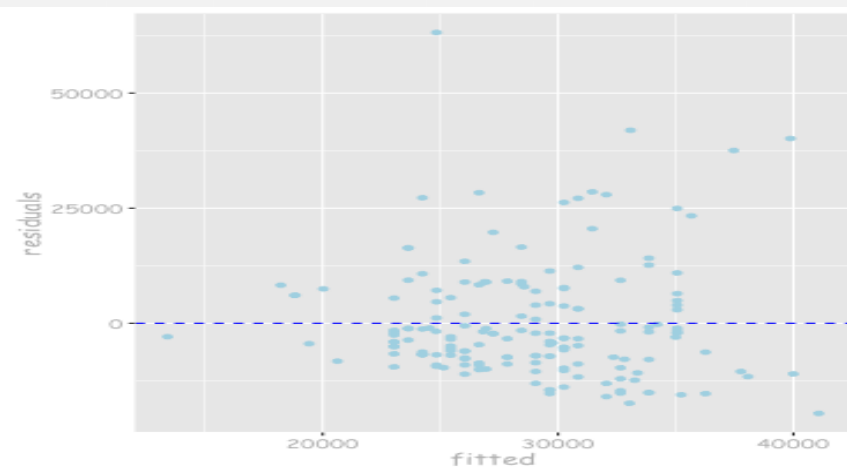
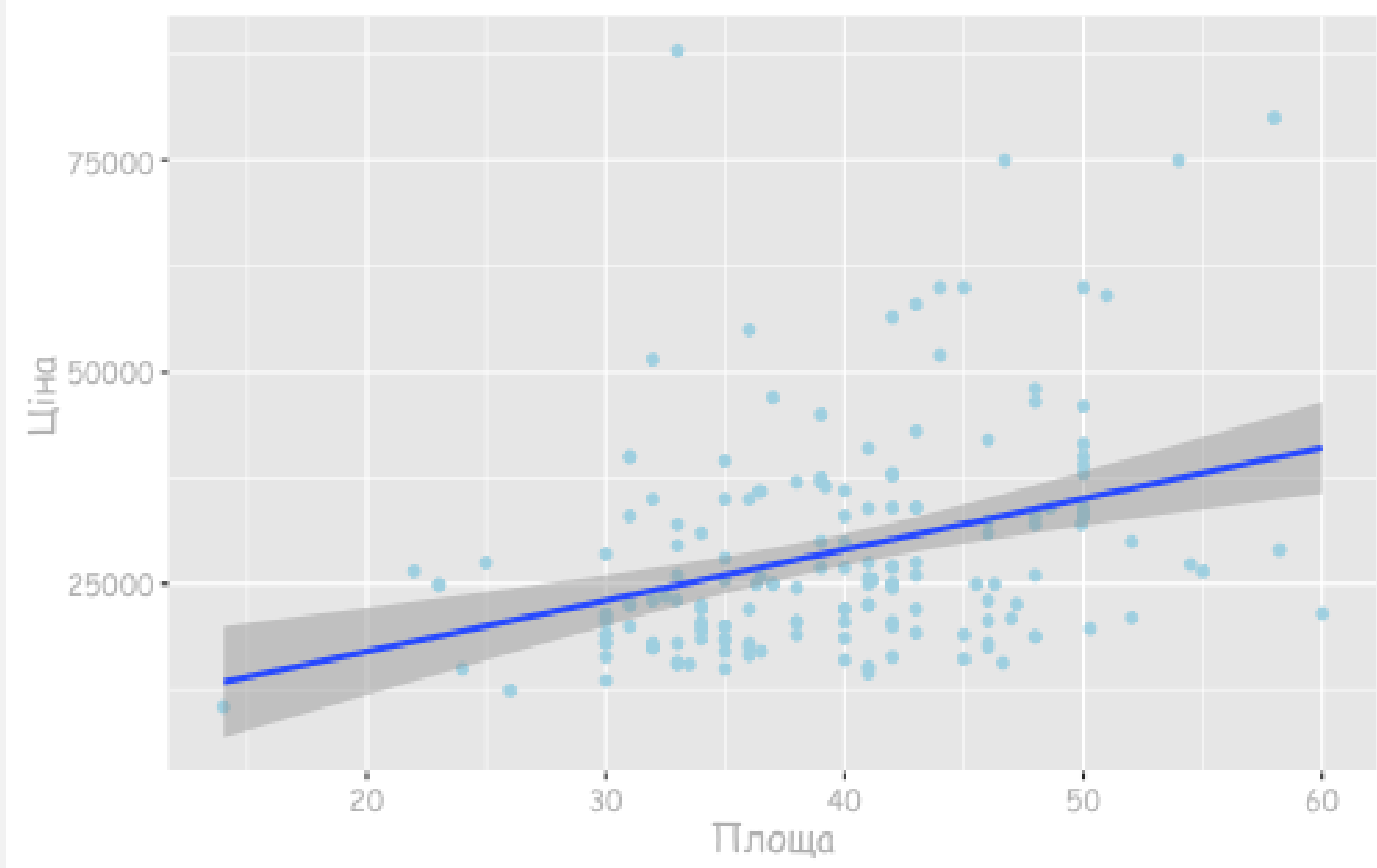
[Check out other apps](#)

[Want to learn more for free?](#)

Regression Model
($R = 0.2267$, $R\text{-squared} = 0.0514$)



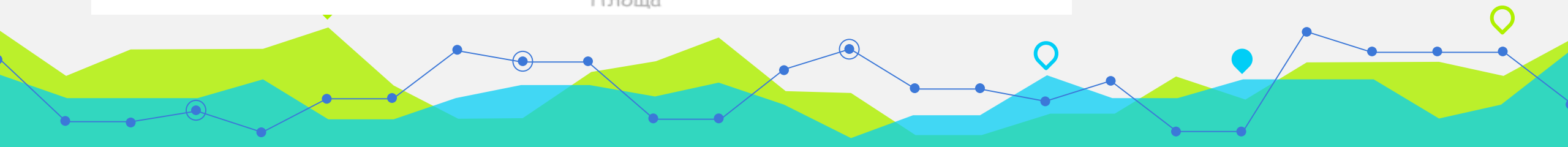
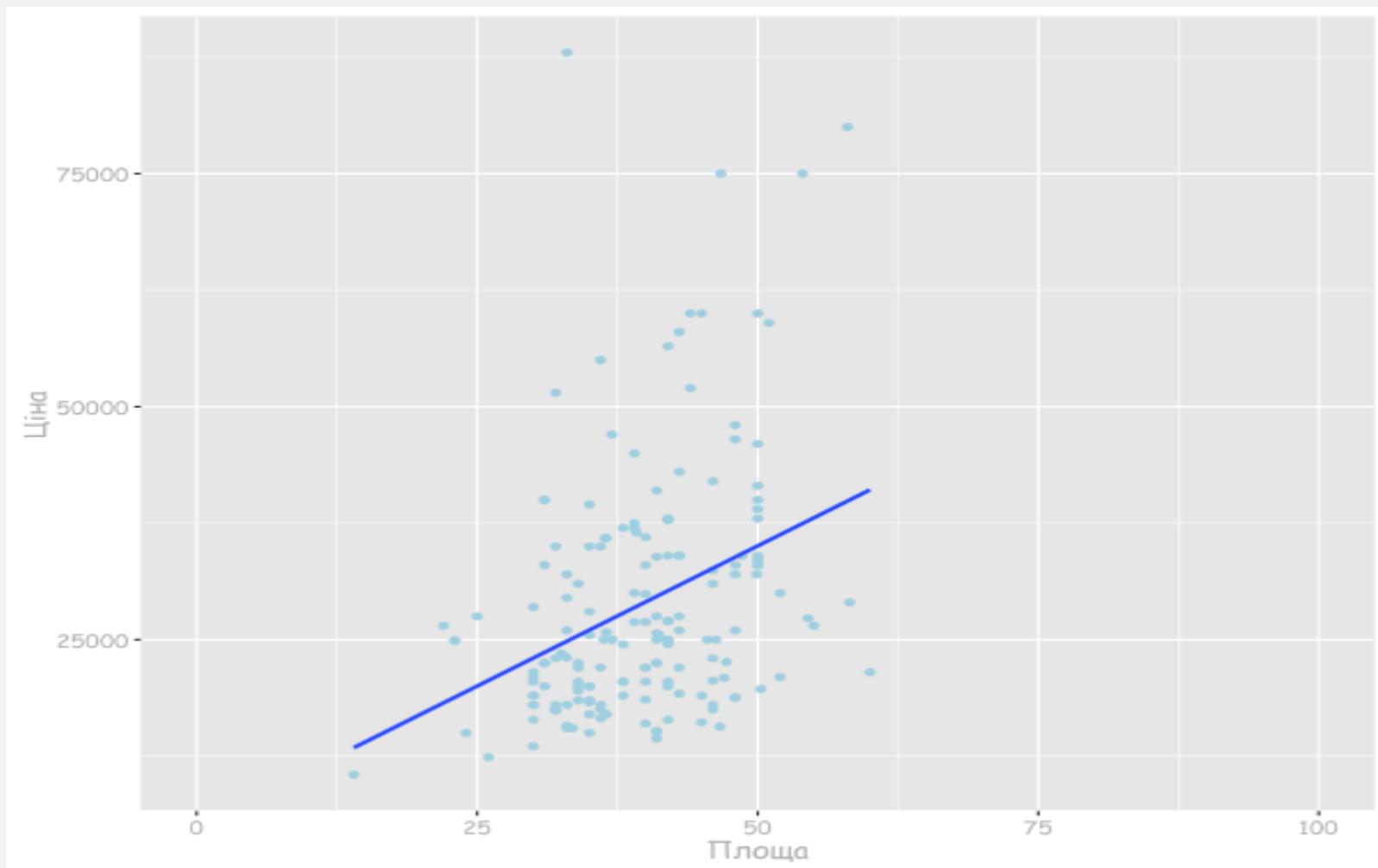






лінійна регресія застереження

екстраполяція



кореляція та причинно-наслідковий зв'язок

