Лекція 2

Вивідна статистика.

Збіжність емпіричних характеристик до теоретичних.

.

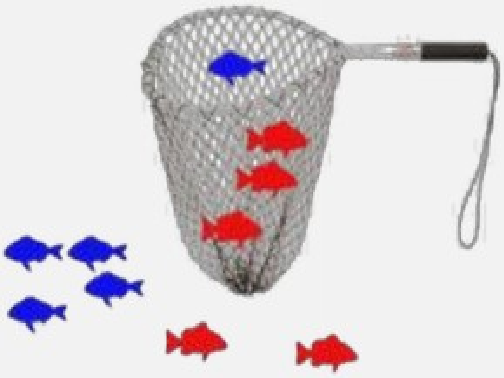
К.ф.-м.н. Щестюк Н.Ю.

Описова статистика цікавиться виключно властивостями спостережуваних даних.

Вивідна статистика дозволяє робити висновки про генеральну сукупність на основі вибірки. Впевненість у цих висновках можна представити чисельно.

Розуміння термінів "генеральна сукупність" та "вибірка" є надзвичайно важливим для розуміння вивідної статистики.

Вибірковий метод. Генеральна сукупність



Генеральна сукупність – усі об'єкти, які хотів би вивчати дослідник при необмеженій кількості ресурсів.

Як сформувати вибірку

Простий випадковий вибір
Всі об'єкти мають однакову можливість бути вибраними. Випадковим чином обирається n об'єктів

Вибір з заміною
Після того, як об'єкт вибрано, він повертається і може бути обраний повторно

Вибір без заміни
Після того, як об'єкт вибрано, він вилучається і не може бути обраний повторно

Стратометричний вибір
Сукупність ділиться на гомогенні групи (населення за рівнем освіти чи віковою групою)

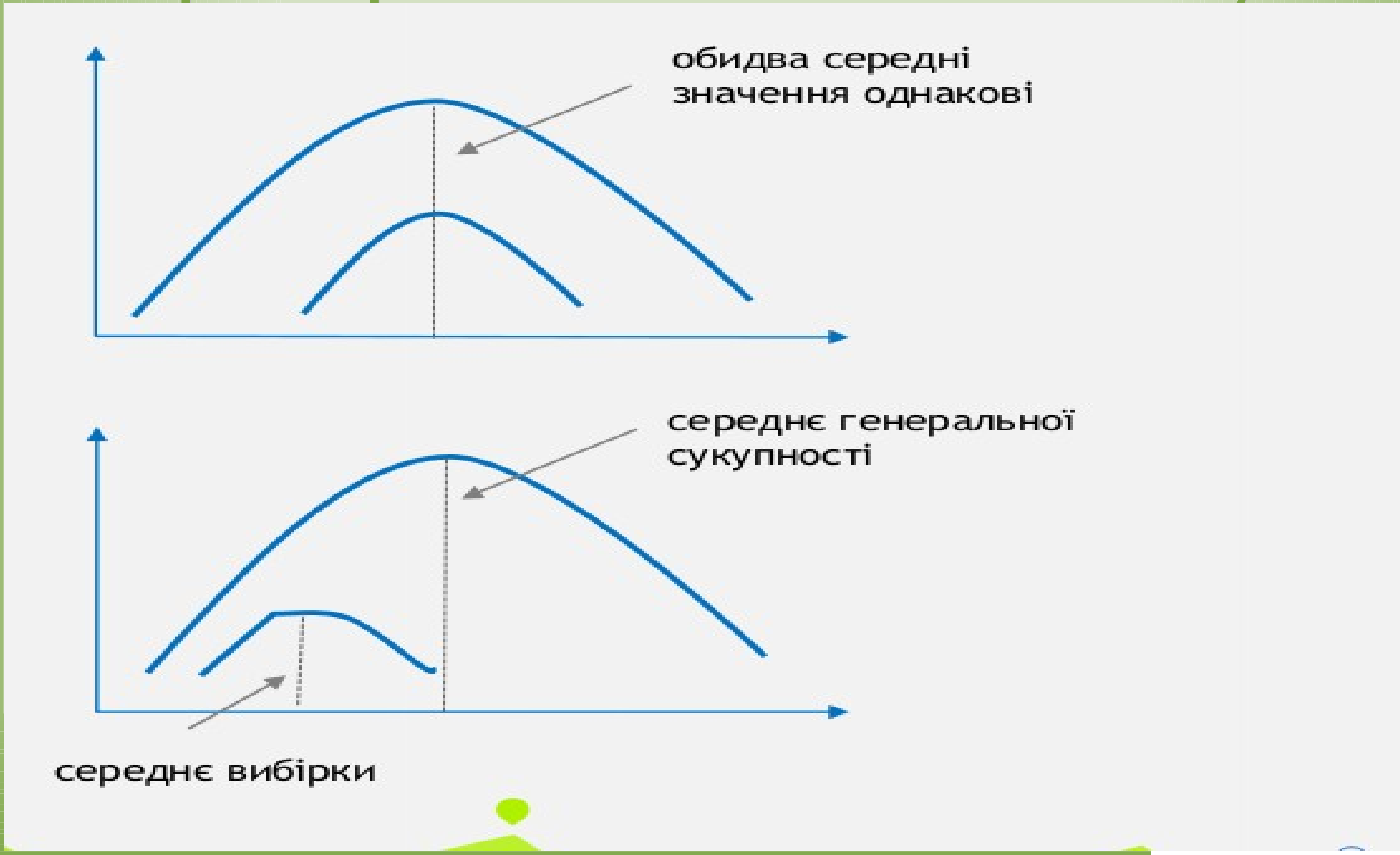
Кластерний вибір
Сукупність ділиться на кластери (місто на райони)

Систематичний вибір
Елементи сукупності впорядковуються і вибирається кожен k-ий елемент (елементи на конвейєрі з метою виявлення дефектів)

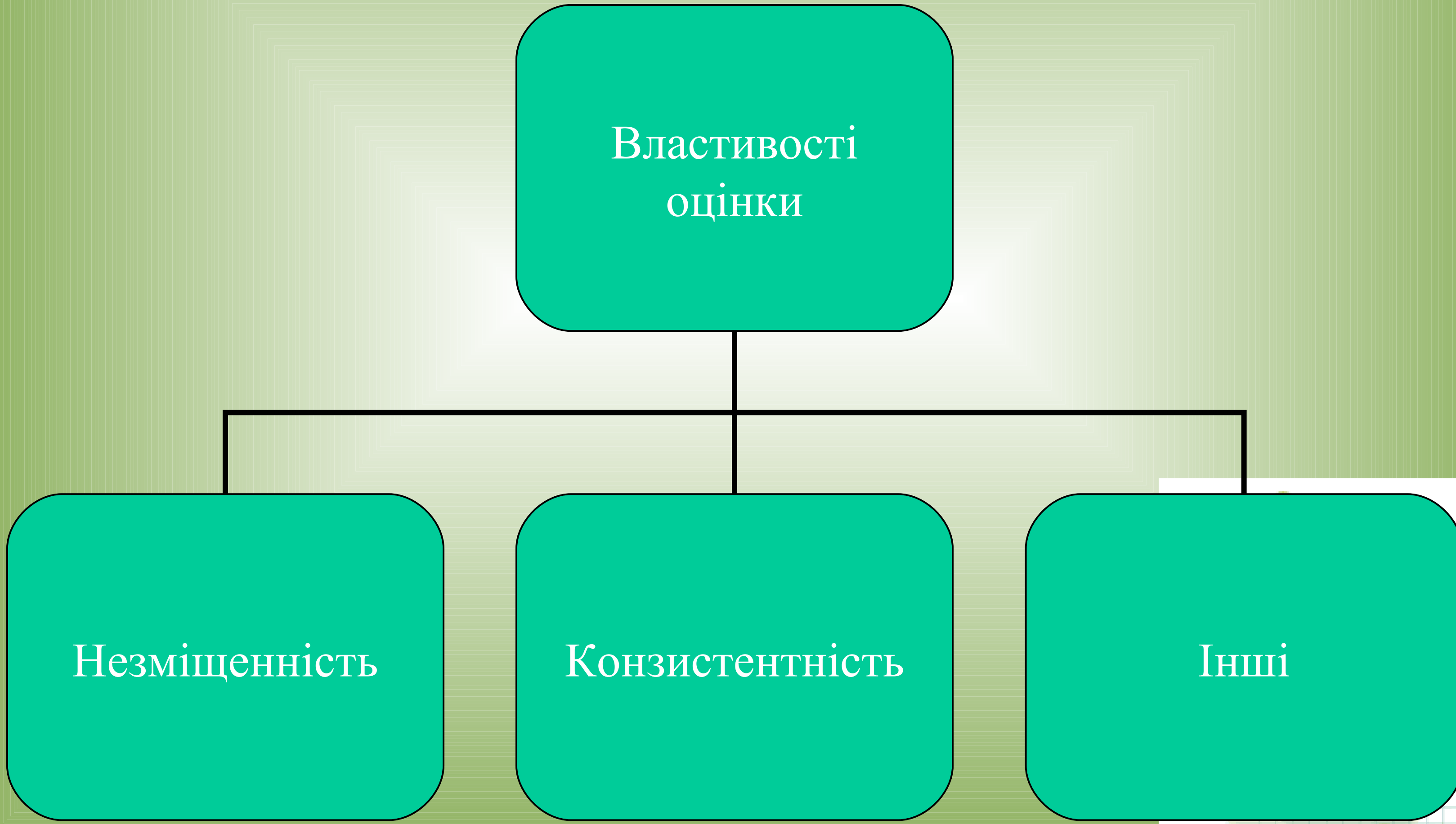
Емпіричні характеристики вибірки: середнє, дисперсія, емпірична функція розподілу

	вибірка	генеральна сукупність
розмір	n	N
середнє значення	$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{n}$
дисперсія	$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$	$\sigma^2 = \frac{\sum (x - \bar{\mu})^2}{N}$
середньоквадратичне відхилення	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
пропорція	$\bar{p} = \frac{n \text{ успіхів}}{n \text{ випробувань}}$	$p = \frac{N \text{ успіхів}}{N \text{ випробувань}}$

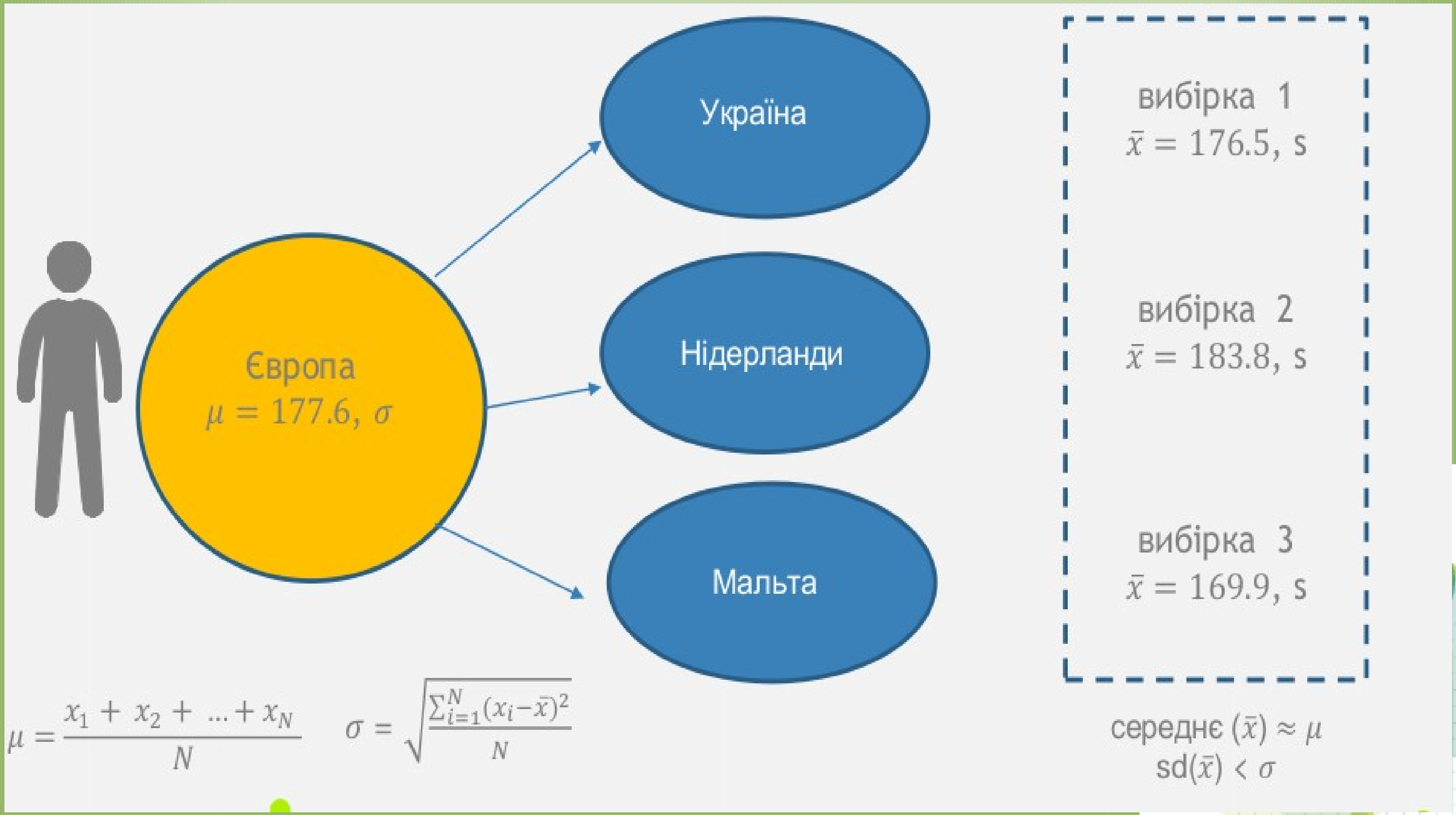
Середнє вибірки і середнє генеральної сукупності (похибка репрезентативності)



Властивості якості емпіричних оцінок



Середнє сукупності вибірок



ЗВЧ у формі Чебишова та Бернуллі

$$P\left(\left|\frac{\sum_{j=1}^{n} z_j}{n} - \frac{\sum_{j=1}^{n} Ez_j}{n}\right| < \varepsilon\right) \geq 1 - \frac{D(z_j)}{n\,\varepsilon^2} \longrightarrow 1$$

$$P\left\{\left|\frac{m}{n} - p\right| > \varepsilon\right\} \leq \frac{pq}{n\varepsilon^2}$$

Припущення щодо вибірки

Генеральна сукупність — це множина всіх значень, яких може набувати дана випадкова величина.

$\left\{ x_1, \ldots, x_n \right\}$ – це **вибірка** спостережень за випадковою величиною ξ із генеральної сукупності (*набір з незалежних і однаково розподілених випадкових величин* ("*копій*")).

$$Ex_j = \mu$$

Незміщенність та консистентність для вибіркового середнього

$$1)\quad E\left(\overline{x}\right)=E\left(\frac{\sum_{j=1}^{n}x_j}{n}\right)=\frac{1}{n}\sum_{j=1}^{n}Ex_j=\frac{1}{n}\sum_{j=1}^{n}\mu=\mu$$

$$2)\ P\left(\left|\frac{\sum_{j=1}^{n}x_j}{n}-\frac{\sum_{j=1}^{n}Ex_j}{n}\right|<\varepsilon\right)\geq 1-\frac{D(x_j)}{n\,\varepsilon^2}\longrightarrow 1$$

Незміщенність (зміщенність) для вибіркової дисперсії

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

1) $$E\left(s^2\right) = E\left(\overline{x^2} - \left(\bar{x}\right)^2\right) = E\left(\overline{x_j^{\,2}}\right) - E\left(\bar{x}\right)^2 =$$

$$= E\left(\overline{x_j^{\,2}}\right) - \left(D\left(\bar{x}\right) + \left(E\bar{x}\right)^2\right) =$$

$$= E\left(x_j^{\,2}\right) - \left(D\left(\frac{\sum x_j}{n}\right)\right) - \left(Ex\right)^2 =$$

$$= D(x) - \frac{n\,D(x)}{n^2} = D(x)\left(1 - \frac{1}{n}\right) = \frac{n-1}{n}D(x)$$

Консистентність для вибіркової дисперсії

$$\text{Нехай } z_j = s_j^2 = x_j^2 - \left(\bar{x}\right)^2 \quad \text{ЗВЧ}: \; P\left(\left|\frac{\sum_{j=1}^{n} z_j}{n} - \frac{\sum_{j=1}^{n} Ez_j}{n}\right| < \varepsilon\right) \ge 1 - \frac{D(z_j)}{n\,\varepsilon^2} \longrightarrow 1$$

$$\frac{\sum_{j=1}^{n} z_j}{n} = \frac{\sum_{j=1}^{n}\left(x_j^2 - \left(\bar{x}\right)^2\right)}{n} = \frac{\sum_{j=1}^{n} x_j^2}{n} - \frac{\sum_{j=1}^{n}\left(\bar{x}\right)^2}{n} = \overline{x^2} - \left(\bar{x}\right)^2$$

$$\frac{\sum_{j=1}^{n} Ez_j}{n} = \frac{\sum_{j=1}^{n} Es_j^2}{n} = \frac{\sum_{j=1}^{n} E\left(x_j^2 - \left(\bar{x}\right)^2\right)}{n} = \frac{nE\left(x_j^2 - \left(\bar{x}\right)^2\right)}{n} = \frac{n-1}{n} D(x_j)$$

Незміщенність для емпіричної функції розподілу

$$F_n^*(y) = \frac{1}{n}\sum_{i=1}^{n} I(x_i < y)$$

$$1)\quad E\left(F^*(y)\right) = \frac{\sum_{j=1}^{n} E(x_j < y)}{n} = \frac{1}{n} n E(x_j < y) =$$

$$= 1P(x_j < y) + 0P(x_j \geq y) = P(x_j < y) = F(y)$$

Консистентність для емпіричної функції розподілу

$$F_n^*(y)=\frac{1}{n}\sum_{i=1}^{n} I(x_i<y)$$

$$\text{Нехай } z_j = I(x_j<y) \qquad \text{ЗВЧ}: \; P\left(\left|\frac{\sum_{j=1}^{n} z_j}{n}-\frac{\sum_{j=1}^{n} Ez_j}{n}\right|<\varepsilon\right)\geq 1-\frac{D(z_j)}{n\varepsilon^2}\longrightarrow 1$$

$$\frac{\sum_{j=1}^{n} z_j}{n}=\frac{\sum_{j=1}^{n} I(x_j<y)}{n}=F^*(y)$$

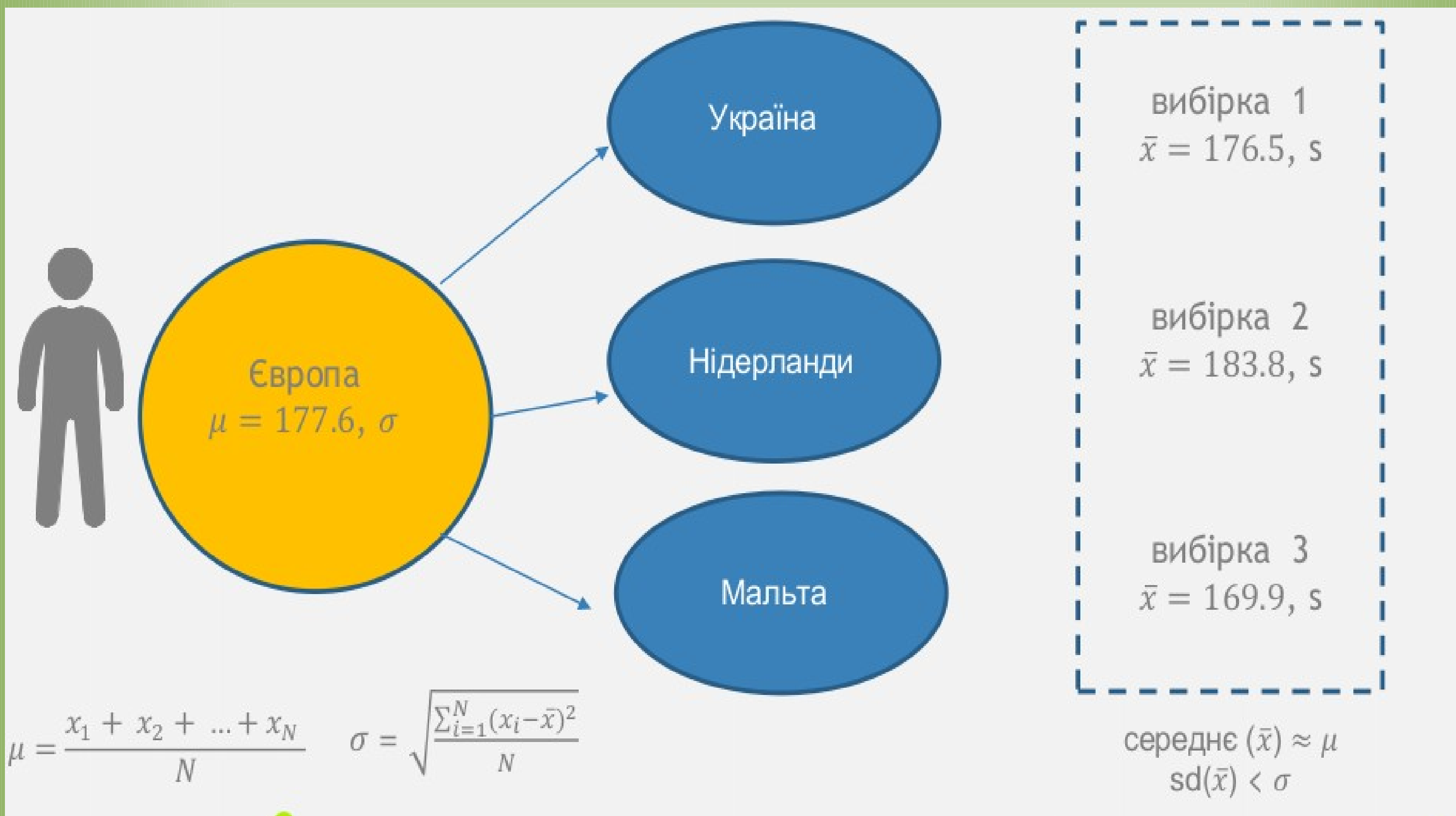
$$\frac{\sum_{j=1}^{n} Ez_j}{n}=\frac{\sum_{j=1}^{n} EI(x_j<y)}{n}=\frac{nEI(x_j<y)}{n}=P(x_j<y)=F(y)$$

ЦГТ і її застосування

$$1)\ P\left(\left|\frac{\sum_{j=1}^{n} z_j}{n} - \frac{\sum_{j=1}^{n} Ez_j}{n}\right| < \varepsilon\right) \longrightarrow \Phi\left(\frac{\bar{z} - \mu}{\sigma}\right)$$

$$2)\ P\left(\left|\frac{m}{n} - p\right| < \Delta\right) = 2\Phi\left(\Delta\sqrt{\frac{n}{pq}}\right)$$

Середнє сукупності вибірок



Довірча ймовірність (confidence level) та розмір вибірки для оцінювання середнього

$$P\left(\left|\bar{x}-\mu\right|<\Delta\right)=\gamma - \textit{довірча ймовірність (рівень довіри)}$$

$$\Delta - \textit{гранична помилка (межа похибки)}$$

$$D\left(\bar{x}\right)=\frac{1}{n^2}\left(D(x_1)+\ldots+D(x_n)\right)=\frac{s^2}{n} - \textit{дисперсія середнього}$$

$$\sigma=\frac{s}{\sqrt{n}} - \textit{стандартна похибка середнього}$$

$$t=\frac{\bar{x}-\mu}{\sigma}=\frac{\Delta\sqrt{n}}{s} - \textit{квантиль нормального розподілу порядку } P$$

$$n=\left(\frac{t\,s}{\Delta}\right)^2 - \textit{розмір вибірки}$$

Довірча ймовірність та розмір вибірки для оцінювання пропорції (успіху)

$$P\left(\left|\frac{m}{n} - p\right| < \Delta\right) = \gamma - \textit{довірча ймовірність (рівень довіри)}$$

Δ - *гранична помилка (межа похибки), виражена в частках одиниці*

$$t = \Delta\sqrt{\frac{n}{pq}} - \textit{квантиль нормального розподілу}$$

порядку P

$$n = \left(\frac{t\sqrt{pq}}{\Delta}\right)^2 - \textit{розмір вибірки}$$

Приклад 1

- Якого розміру має бути вибірка, щоб оцінити середній обєм випитого за місяць пива для людей певного регіону. Випадкова похибка має не перевищувати 0,1 л з довірчою ймовірністю 95%. Відомо, що s=2л

$$t = 1{,}96 - \textit{квантиль нормального розподілу}$$

$$\textit{порядку } P = 0.95$$

$$n = \left(\frac{t\, s}{\Delta} \right)^2 = \left(\frac{1.96 * 2}{0.1} \right)^2 \approx 1600 - \textit{розмір вибірки}$$

Приклад 2

- Якого розміру має бути вибірка, щоб оцінити частку громадян України, які мають намір прийти на вибори. Випадкова похибка має не перевищувати 2%. З попередніх досліджень для частки осіб, які мають намір прийти має не перевищувати 0,9 з довірчою ймовірністю 95%.

$$t = 1{,}96 - \textit{квантиль нормального розподілу}$$

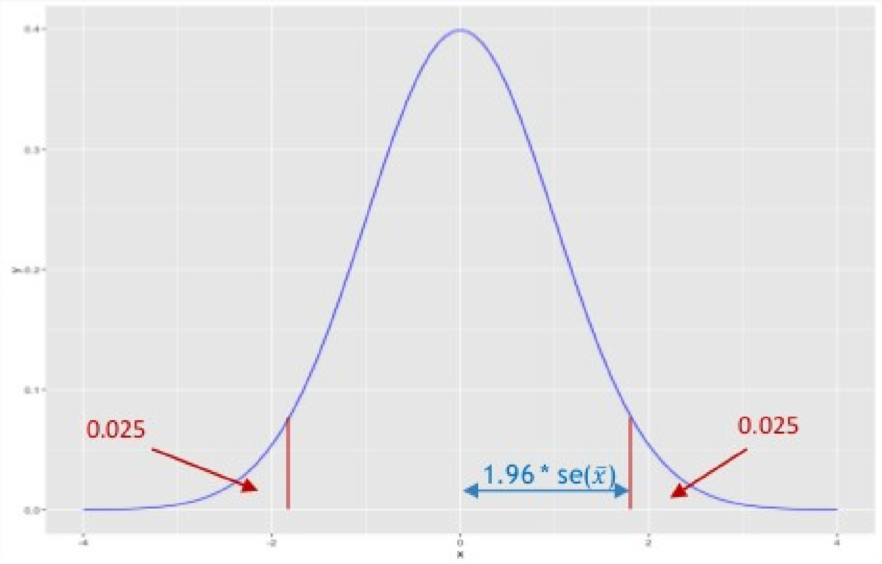
$$\textit{порядку } P = 0.95$$

$$n = \left(\frac{t\sqrt{pq}}{\Delta} \right)^2 = \left(\frac{1.96\sqrt{0.9 * 0.1}}{0.02} \right)^2 \approx 900 - \textit{розмір вибірки}$$

довірчий інтервал для середнього значення

$\bar{x} \pm Z_{95\%} * se(\bar{x})$, де
$se(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

$\bar{x} \pm 1.96 * se(\bar{x})$, де
$se(\bar{x}) = \frac{\sigma}{\sqrt{n}}$



$\mu_{\bar{x}} = \mu$
$se(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

1.96 * se($\bar{x}$) – межа по
для рівня довіри 95%

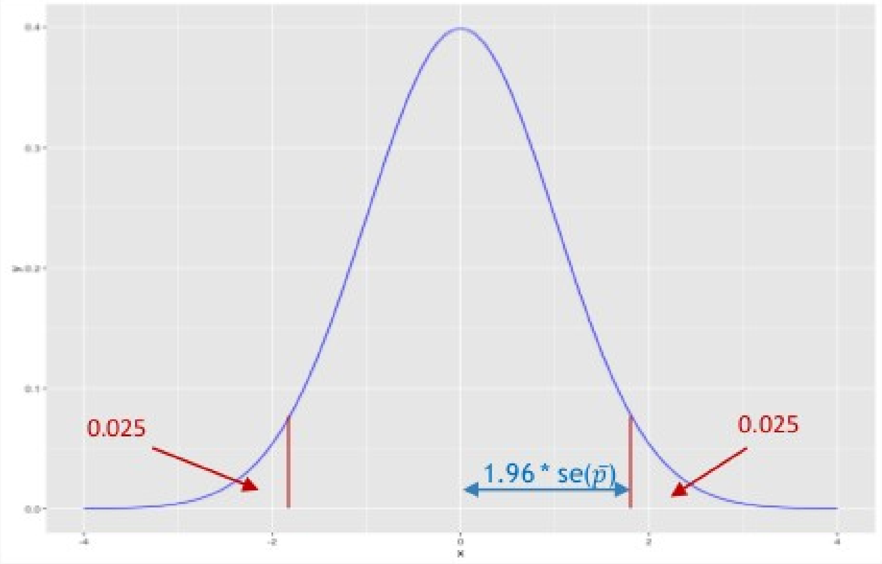
довірчий інтервал для пропорції

$p \pm Z_{95\%} * \text{se}(\bar{p})$, де

$\text{se}(\bar{p}) = \sqrt{\frac{p(1-p)}{n}}$

$p \pm 1.96 * \text{se}(\bar{p})$, де

$\text{se}(\bar{p}) = \sqrt{\frac{p(1-p)}{n}}$



$\mu_{\bar{p}} = p$,

$\text{se}(\bar{p}) = \sqrt{\frac{p(1-p)}{n}}$

1.96 * se($\bar{p}$) – межа пох
для рівня довіри 95%

Приклад

довірчий інтервал для пропорції

Серед 935 випадковим чином обраних респондентів на питання “чи вирі́те ви в існування розумного життя на інших планетах?” ствердно відповіли 60%

$\bar{p} = 0.6, n = 935$

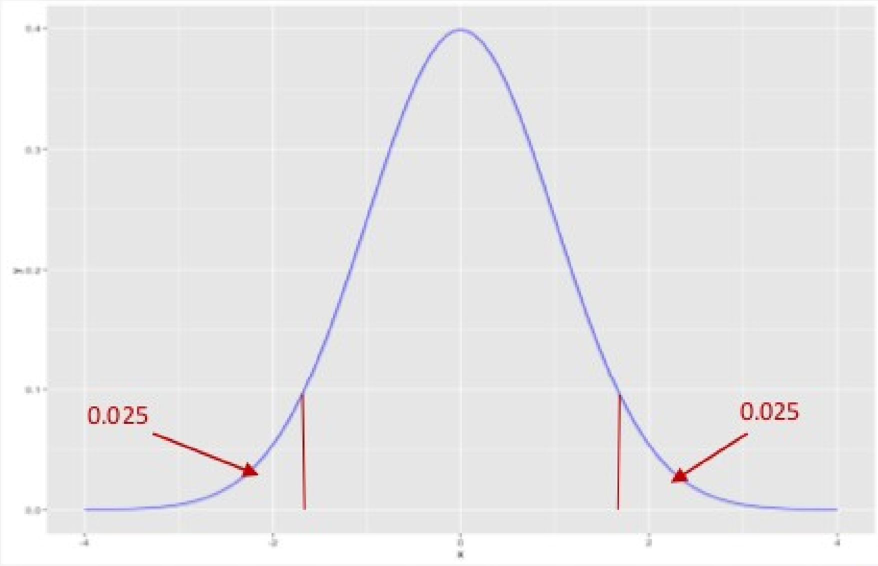
$p \pm Z_{95\%} * se(\bar{p})$, де $se(\bar{p}) = \sqrt{\frac{p(1-p)}{n}}$

$0.6 \pm 1.96 * se(\bar{p})$, де $se(\bar{p}) = \sqrt{\frac{0.6(1-0.6)}{935}} = 0.016$

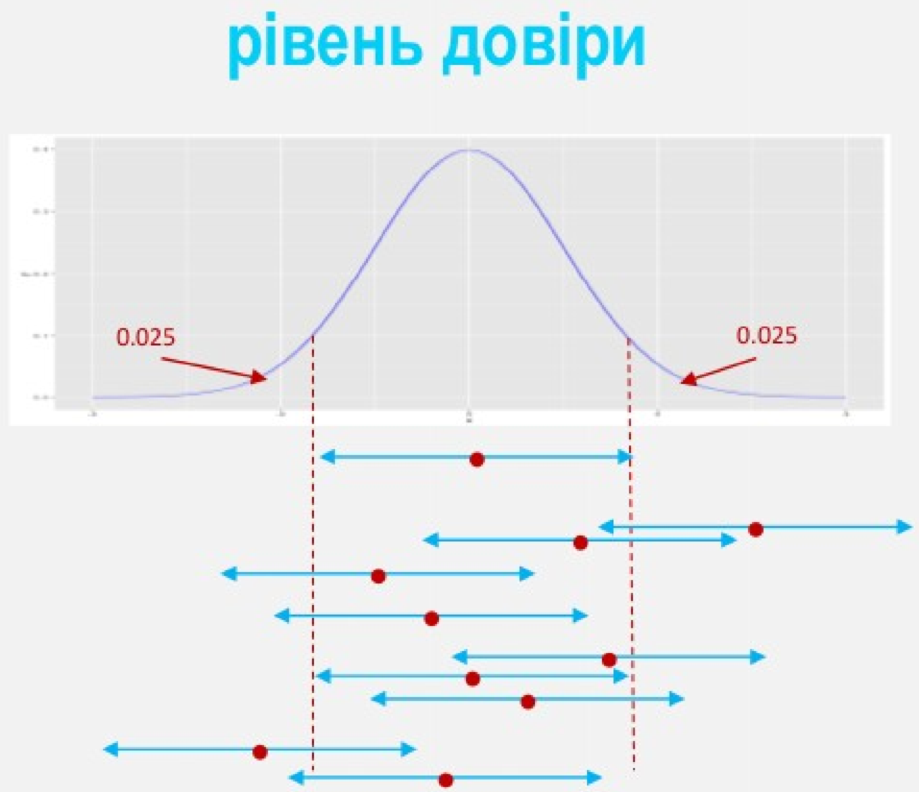
$0.6 \pm 1.96 * 0.016$

0.6 ± 0.03136

$[0.56864, 0.63136]$



Довірчий інтервал



Дякую за увагу!