Аналіз даних та статистичне виведення

Тиждень 4

вивідна статистика

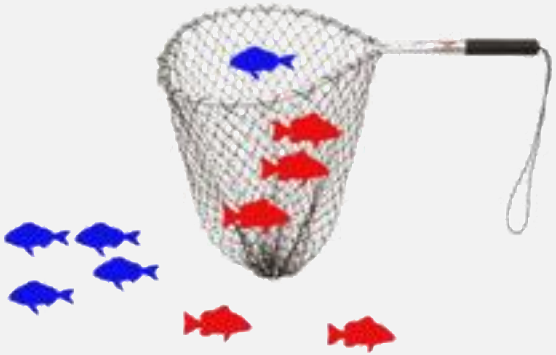
описова та вивідна статистика

Описова статистика – вивчає властивості спостережуваних даних.

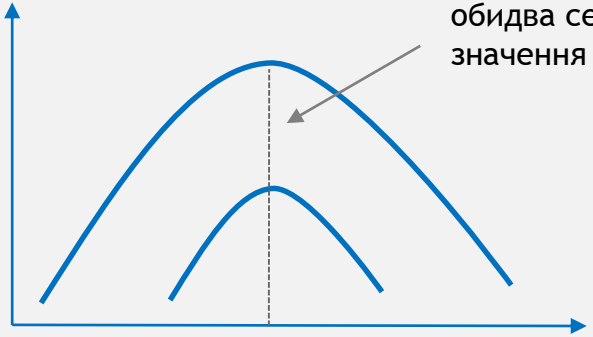
Вивідна статистика – виводимо припущення про властивості розподілу даних з яких походять спостережувані дані.

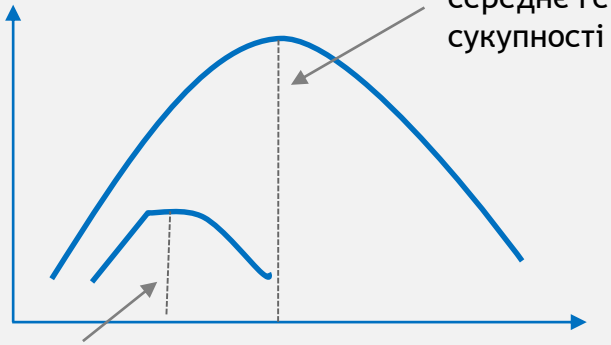
	вибірка	генеральна сукупність
розмір	n	N
середнє значення	$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{n}$
дисперсія	$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$	$\sigma^2 = \frac{\sum (x - \bar{\mu})^2}{N}$
середньоквадратичне відхилення	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
пропорція	$\bar{p}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

вибірка



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





Репрезентативна вибірка

Представляє генеральну сукупність, можна використовувати для вивідної статистики

Нерепрезентативна вибірка

Вибірка та генеральну сукупність мають різні характеристики. Використання цієї вибірки призведе до неправильних результатів аналізу

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

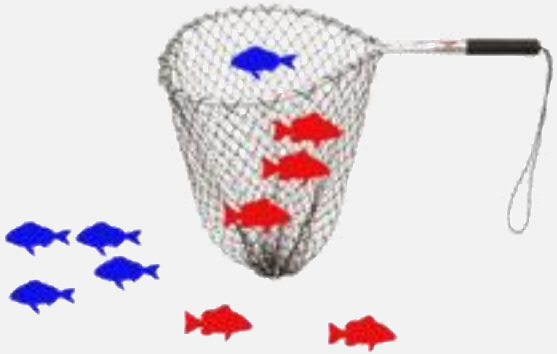
Вибір без заміни
Після того, як об'єкт вибрано, він вилучається і не може бути обраний повторно

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 конвейері з метою виявлення дефектів)

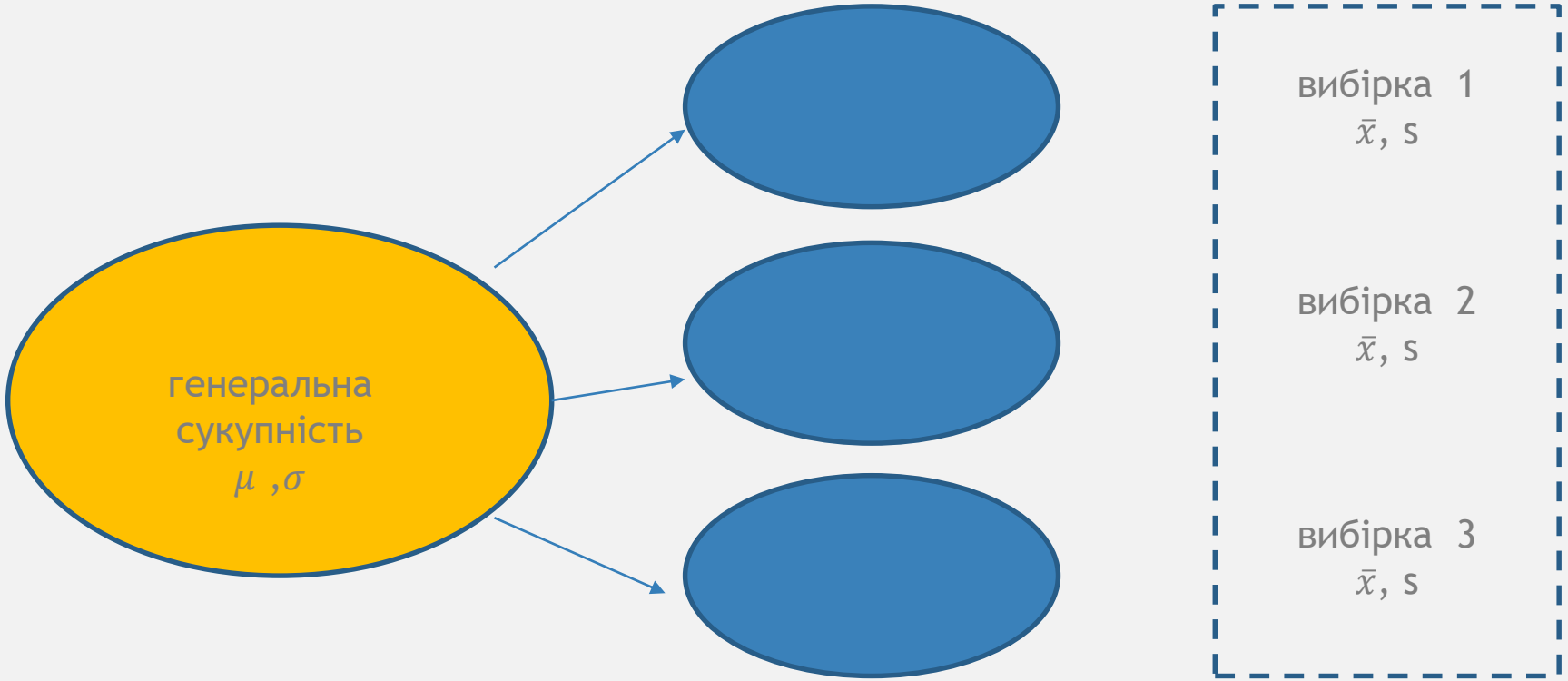
типові помилки при формуванні вибірки



- Помилка недоохоплення
- “Волонтерська” вибірка
- “Помилка тих, хто вижив”

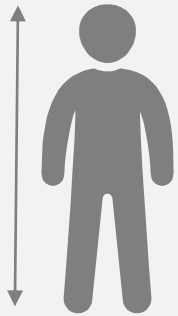
	вибірка	генеральна сукупність
розмір	n	N
середнє значення	$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{n}$
дисперсія	$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$	$\sigma^2 = \frac{\sum (x - \bar{\mu})^2}{N}$
середньоквадратичне відхилення	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
пропорція	$\bar{p}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

централна гранична теорема



розподіл вибірки

вибірковий розподіл



Європа
$\mu = 177.6,\ \sigma$

Україна

Нідерланди

Мальта

вибірка 1
$\bar{x} = 176.5,\ s$

вибірка 2
$\bar{x} = 183.8,\ s$

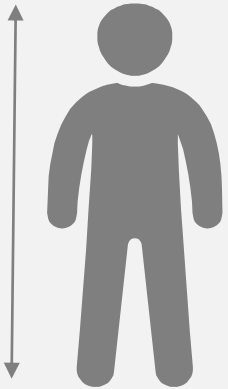
вибірка 3
$\bar{x} = 169.9,\ s$

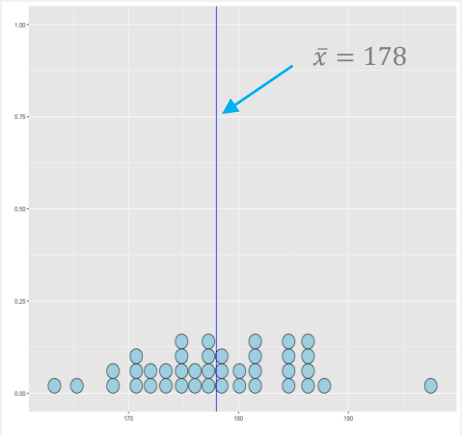
$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

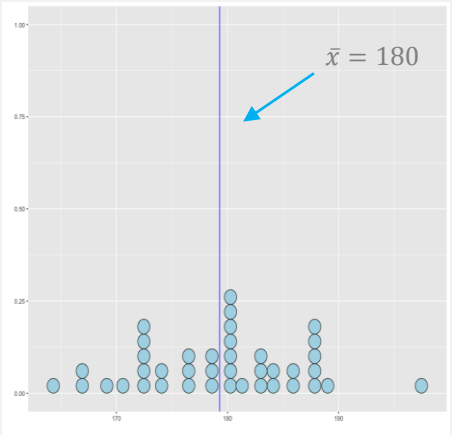
середнє $(\bar{x}) \approx \mu$
$\mathrm{sd}(\bar{x}) < \sigma$

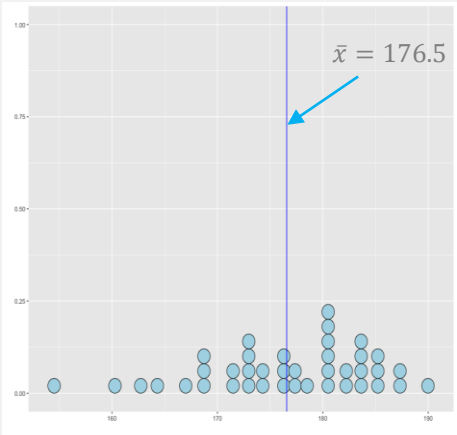
$\bar{x} = 178$ $\bar{x} = 180$ $\bar{x} = 176.5$

?



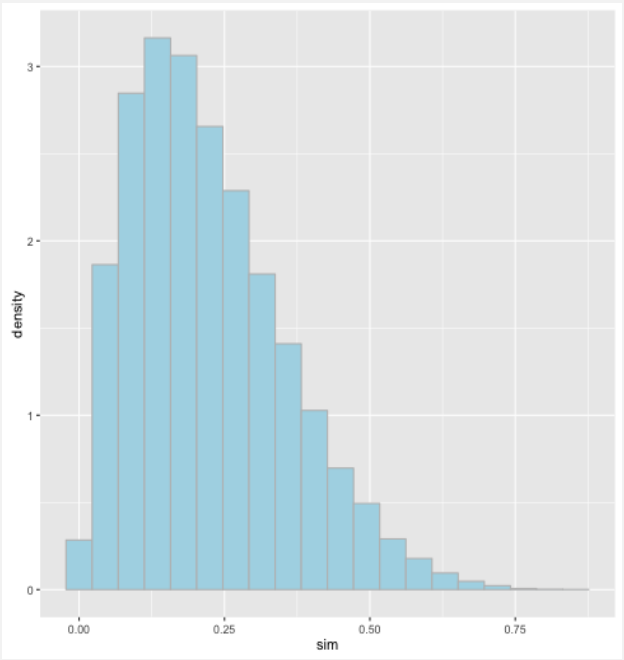


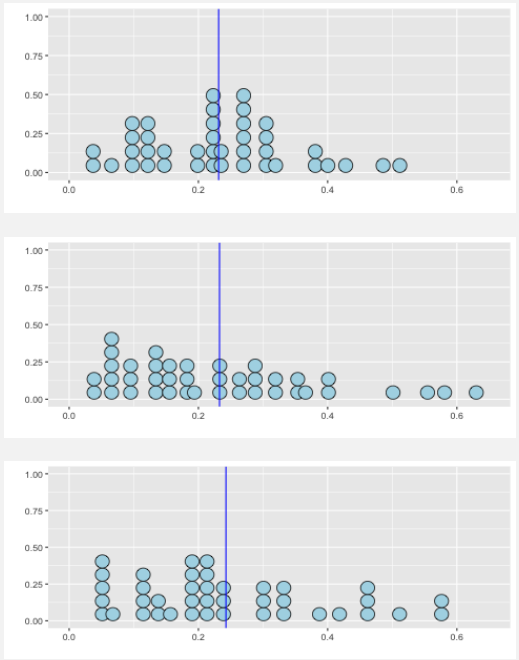


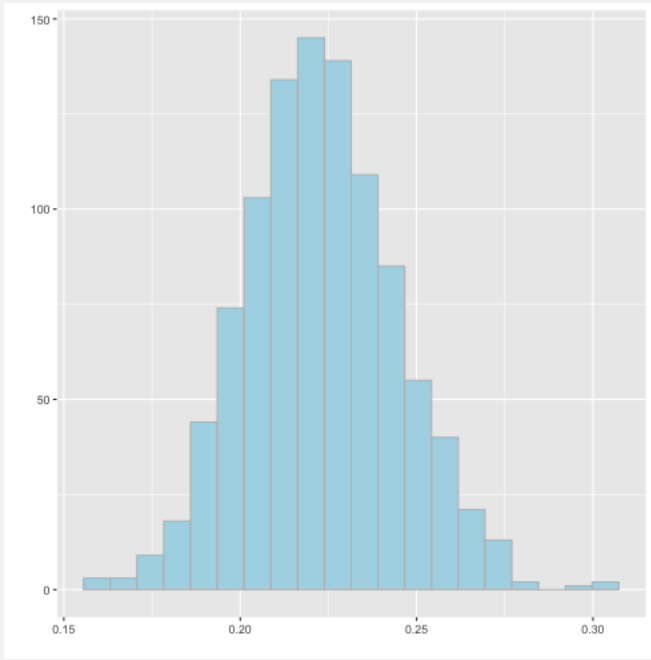


$$\mu \approx \frac{\overline{x_1}+\overline{x_2}+ \ldots+\overline{x_N}}{N} = 177.6$$

три розподіли







https://gallery.shinyapps.io/CLT_mean/

центральна гранична теорема

Неважливо, який розподіл має змінна у популяції (генеральній сукупності)

Вибірковий розподіл середніх значень вибірок має приблизно нормальний розподіл, якщо розмір вибірки принаймі 30

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

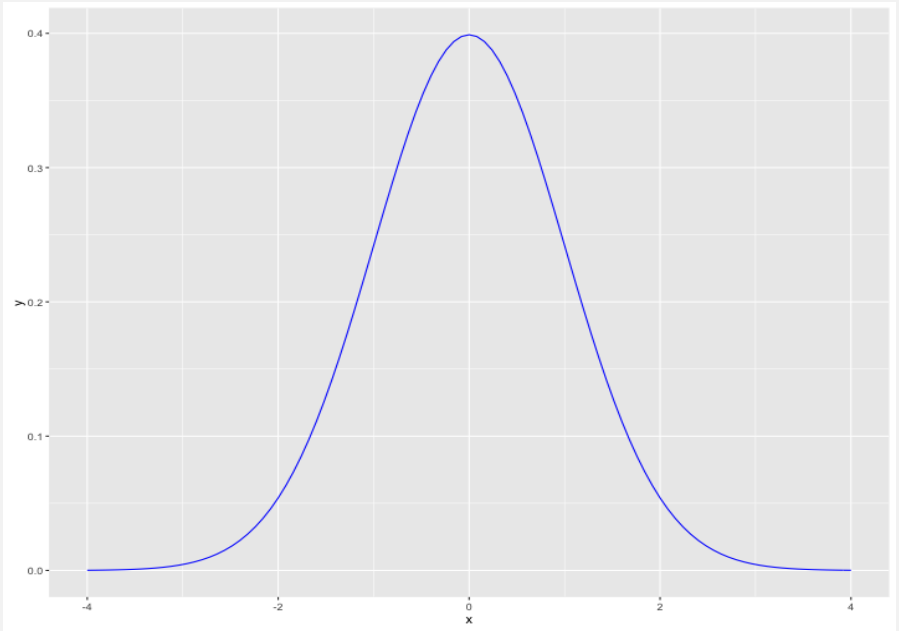
Вибірковий розподіл пропорцій вибірок має приблизно нормальний розподіл за умови наявності принаймі 15 успіхів та 15 невдач

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

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

застосування центральної граничної теореми

Ви збираєтесь на пробіжку, яка триватиме 2 години. Яка ймовірність що плейлист з 40 пісень не закінчиться протягом пробіжки? (середня довжина пісні 3.45 хв, середньоквадратичне відхилення – 1.63 хв)



2 години = 120 хв
$P(x_1 + x_2 + \ldots + x_{40}) > 120$
$P(\bar{x} > 3)$?

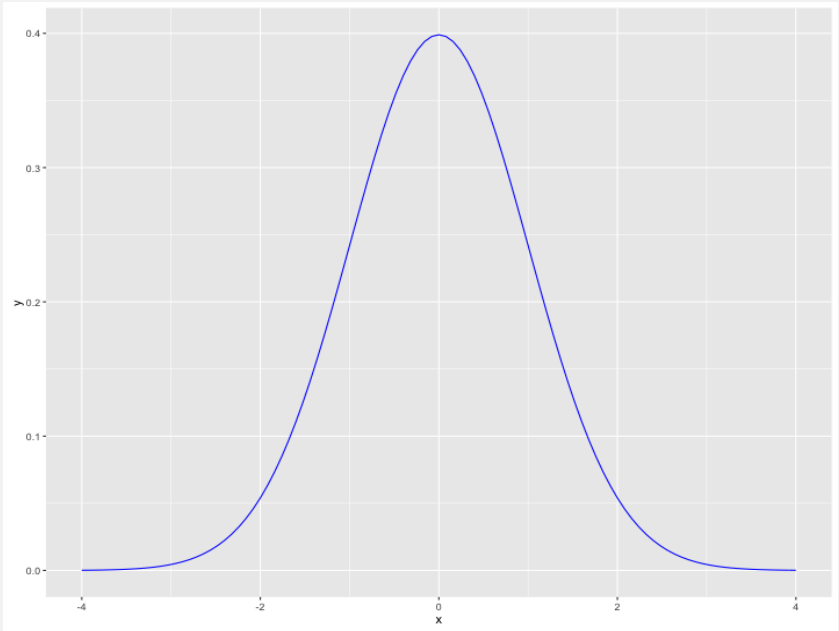
$$se = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{40}} = 0.258$$

$$z = \frac{3 - 3.45}{0.258} = -1.74$$

$P(\bar{x} > 3) = 0.959$

застосування центральної граничної теореми

Припустимо, що частка всіх студентів університету, які вживали енергетики протягом останніх 6 місяців становить p = 0.40. Для вибірки n = 200 студентів, яка ймовірність, що відсоток тих, хто вживав вживав енергетики протягом останніх 6 місяців менша ніж 32%?



$P(\bar{p} < 0.32)$?

np = 200*0.40 = 80, n(1-p)=200*(1-0.40)=120

$se(\bar{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.40(1-0.40)}{200}} = 0.0346$

$z = \frac{0.32-0.40}{0.0346} = -2.31$

$P(\bar{p} < 0.32) = 0.01$

довірчий інтервал для пропорції

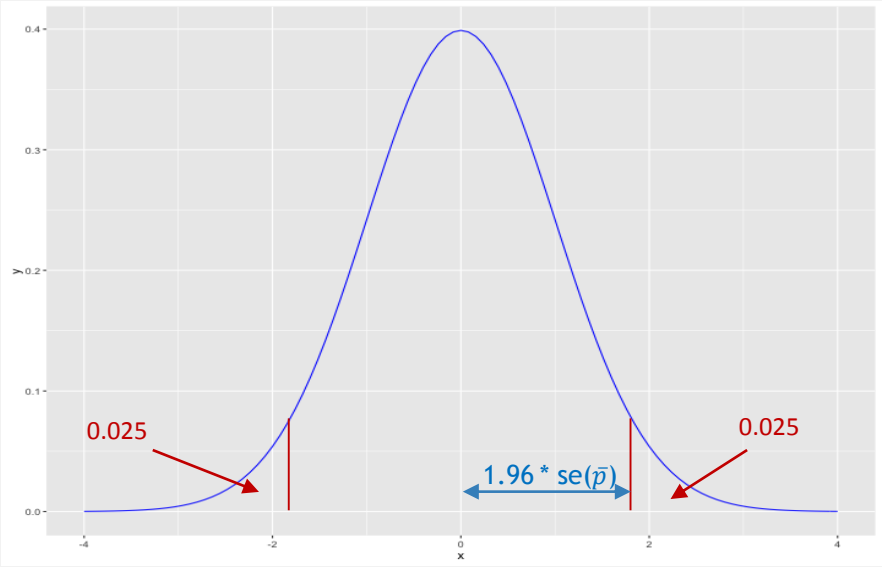
довірчий інтервал для пропорції

$p \pm Z_{95\%} * se(\bar{p})$, де

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

$p \pm 1.96 * se(\bar{p})$, де

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



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

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

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

довірчий інтервал для пропорції

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

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

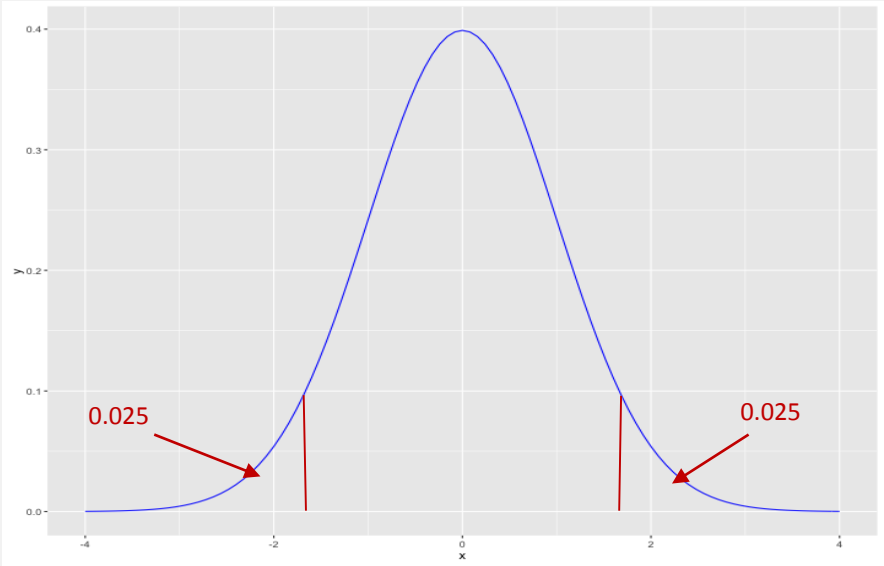
$p \pm Z_{95\%} * se(\bar{p})$, де $se(\bar{p}) = \sqrt{\frac{p(1-p)}{n}}$

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

$0.6 \pm 1.96 * 0.016$

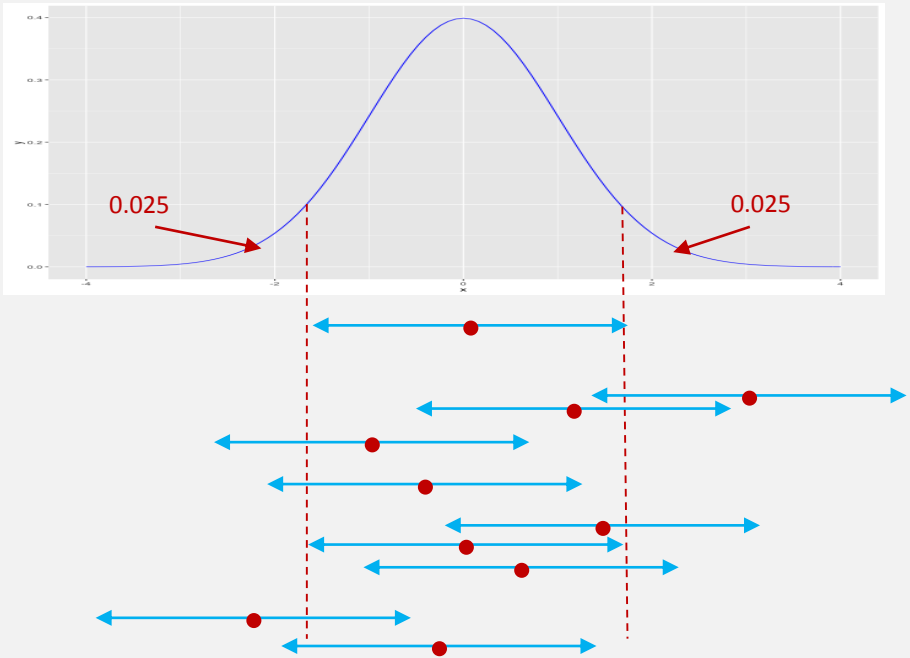
0.6 ± 0.03136

[0.56864, 0.63136]



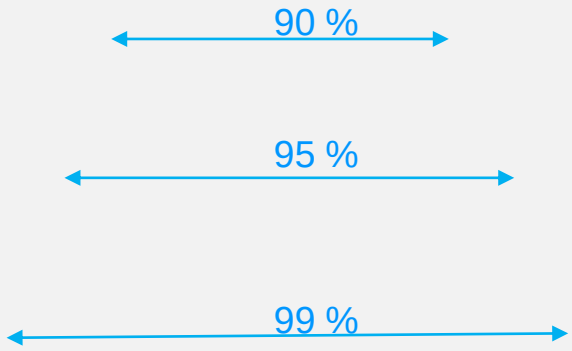
рівень довіри

рівень довіри



рівень довіри

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



90% CI: $p \pm Z_{90\%} * se(\bar{p}) = 0.6 \pm 1.64 * 0.016$, [0.57376, 0.62624]

95% CI: $p \pm Z_{95\%} * se(\bar{p}) = 0.6 \pm 1.96 * 0.016$, [0.56864, 0.63136]

99% CI: $p \pm Z_{99\%} * se(\bar{p}) = 0.6 \pm 2.58 * 0.016$, [0.55872, 0.64128]

розмір вибірки

розмір вибірки

Нехай дослідження показало, що 43% дорослих віком від 25 до 35 років співає в душі. Дослідник хоче визначити, чи це справедливо для дорослих віком від 35 років. Яким має бути розмір вибірки, щоб мати межу похибки рівну 5% для рівня довіри 90%?

$$p \pm Z_{90\%} * se(\bar{p}), \text{ де } se(\bar{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Межа похибки: $Z_{90\%}\sqrt{\frac{p(1-p)}{n}}$, $Z_{90\%}$=1.64485

$$0.05 = 1.64485 * \sqrt{\frac{0.43*0.57}{n}}$$

$$\sqrt{n} = 1.64485*\frac{\sqrt{0.43*0.57}}{0.05} = 16.2865$$

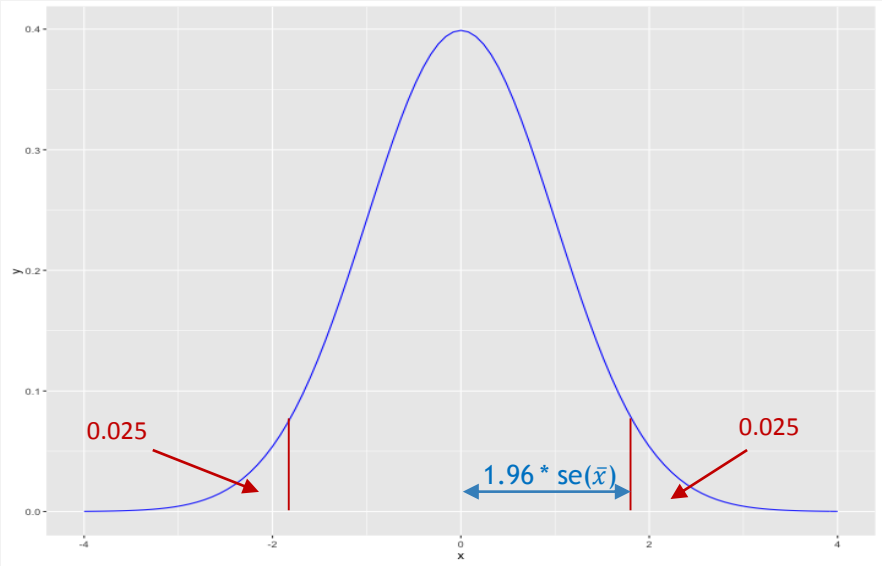
$n = 16.2865^2 = 265.25$, тобто потрібно опитати 266 респондентів щоб отримати межу похибки 5%

довірчий інтервал для середнього значення

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

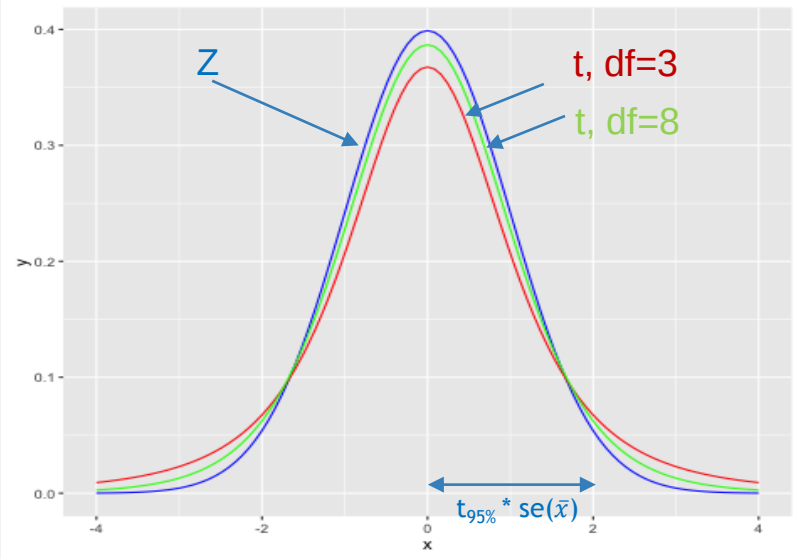


$\mu_{\bar{x}} = \mu$
$se(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

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

довірчий інтервал для середнього значення

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



$\mu_{\bar{x}} = \mu$

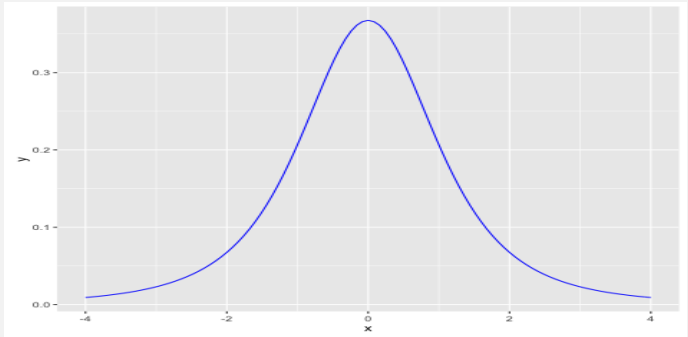
$se(\bar{x}) = \frac{s}{\sqrt{n}}$

df = n - 1

довірчий інтервал для середнього значення

Для визначення екваторіальногр радіусу планети Юпітер провели 40 незалежних вимірювань. Ці вимірювання мають середнє значення $\bar{x}$ = 71492 км та середньоквадратичне відхилення s = 28 км. Знайдіть 90% довірчий інтервал для екваторіального радіуса Юпітера.

$\bar{x} \pm t_{90\%} * se(\bar{x})$, де $se(\bar{x}) = \frac{s}{\sqrt{n}}$



df = n-1 = 39

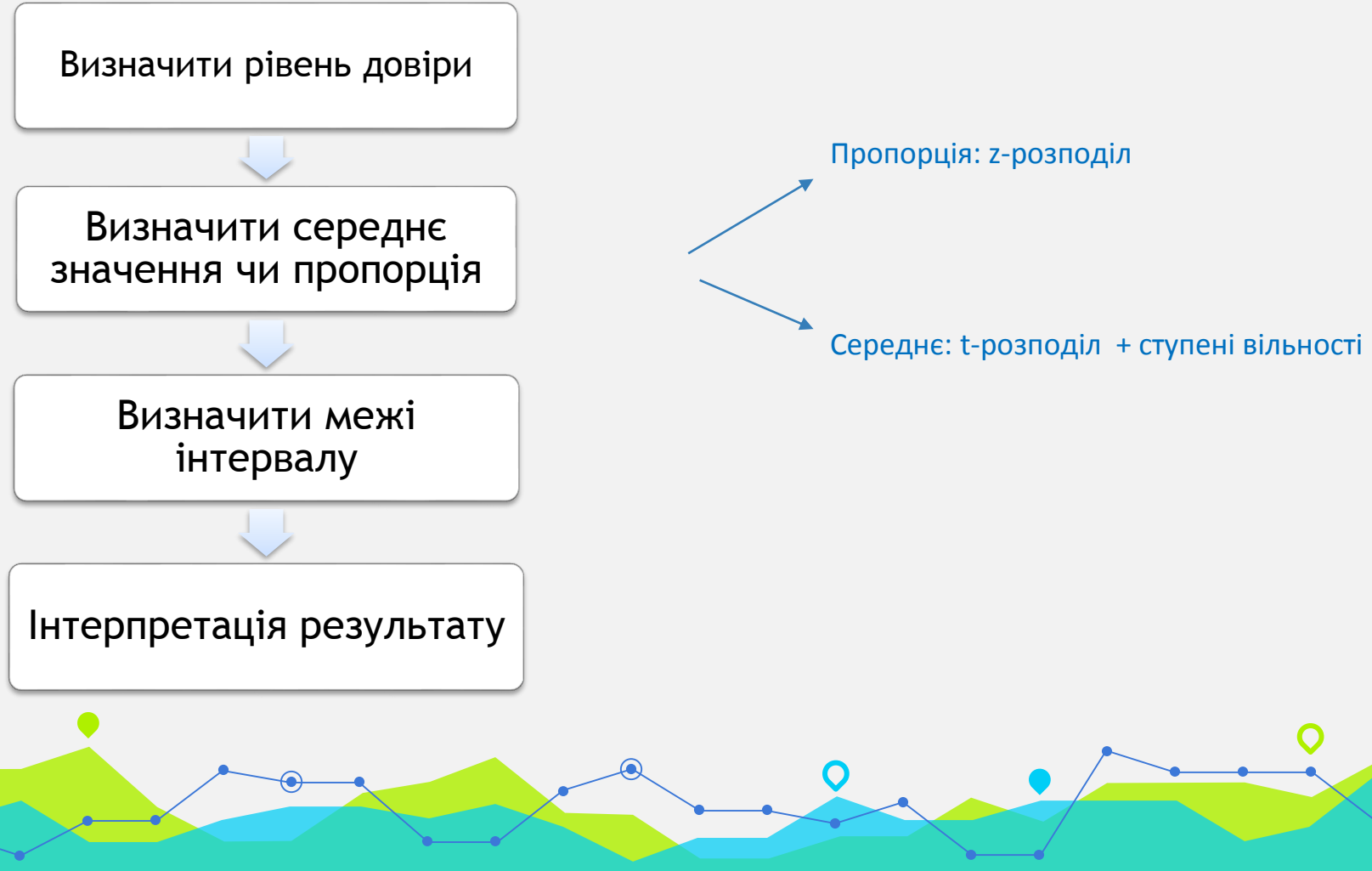
$t_{90\%}$= 1.3, $se(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{28}{\sqrt{40}} = 4.4272$

$\bar{x} \pm 1.3 * 4.4272$

71492 ± 5.755

$[71486.24, 71497.76]$

покроковий план побудови довірчого інтервалу


Визначити рівень довіри
Визначити середнє значення чи пропорція
Визначити межі інтервалу
Інтерпретація результату
Пропорція: z-розподіл
Середнє: t-розподіл + ступені вільності