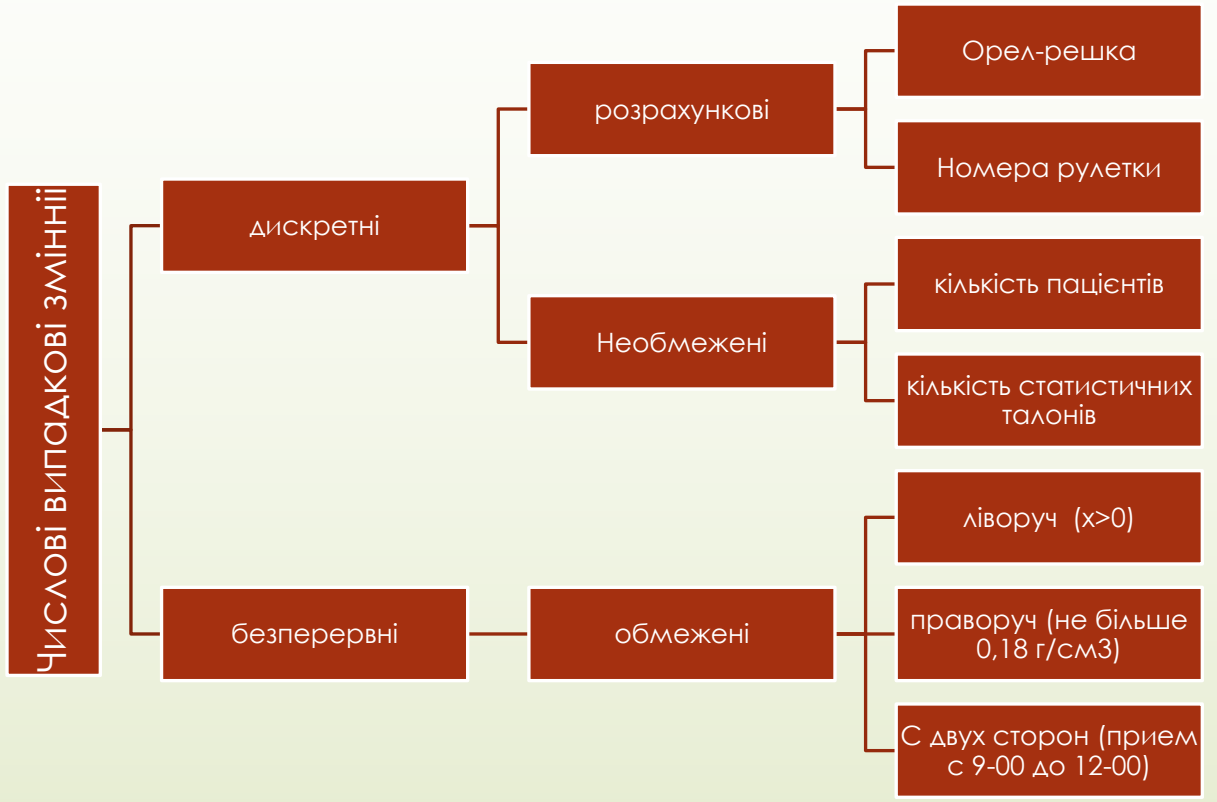
Розподіл та характеристики різних типів розподілу

Основні визначення

Випадкові змінні

- якісні (відсутня упорядкованість)
  - колір авто
- порядкові (після а йде б, і т.п. але а та б неможливо порівнювати на хто більше, хто менше)
  - червоний - жовтий - зелений
- рангові (упорядковані)
  - землетрус 1— 2-…9 балів
    1 менше 9
  - брудно-неприбрано чисто-прибрано
- числові
  - дискретні
    - 0,1,…10 (цілі числа)
  - безперервні
    - 0…10 (будь-яка дріб від та доо)

Випадкові змінні



Закон розподілу

- Найбільш повну, Найповнішою, вичерпної характеристикою випадкової змінної є закон розподілу.
- Закон розподілу - функція (таблиця, графік, формула), що дозволяє визначати ймовірність того, що випадкова змінна X приймає певне значення x_i або потрапляє в певний інтервал.
- Якщо випадкова змінна має даний закон розподілу, то кажуть, що вона розподілена за цим законом або підпорядковується цим законом розподілу.

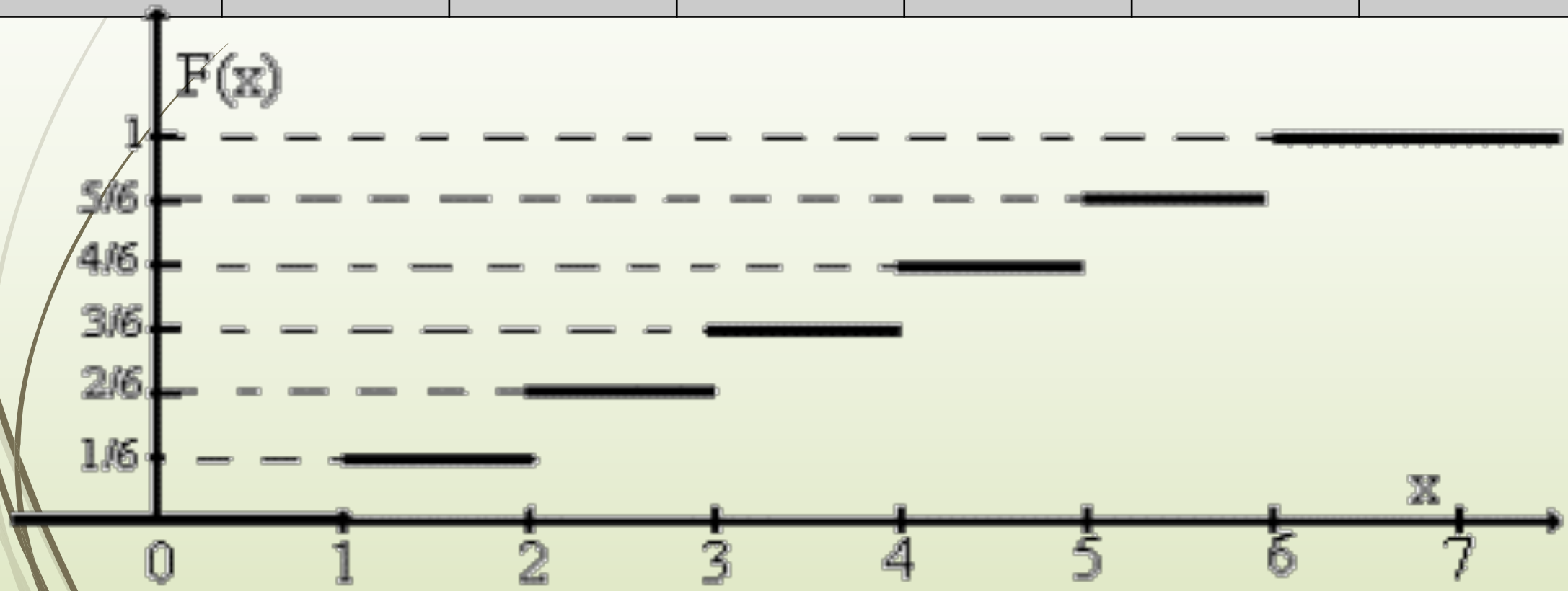
Способи завдання законів розподілу

- Графік (номограмма)
- Таблиця (розподіл Стьюдента)
- Формула $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-a)^2}{2\sigma^2}}$

Функція розподілу

- Імовірність того, що випадкова змінна у випробуванні виявиться менше заданої

Значения x_i:	**1**	**2**	**3**	**4**	**5**	**6**
Вероятности $p(x_i)$	1/6	1/6	1/6	1/6	1/6	1/6



Властивості функцій розподілу

- $F(-\infty) = 0$
- $F(+\infty) = 1$
- $F(x) -$ функція, яка не зменьшуєтьсях

Закон розподілу

- Для безперервних випадкових змінних вводиться поняття щільності розподілу p (x), яка є похідна від функції розподілу (для дискретних - ймовірності можна задати у вигляді таблиці значень).
- $p(x) = \frac{dF(x)}{dx}$
- $F(x) = \int p(x)dx$
- Часто під законом розподілу розуміють саме закон розподілу ймовірностей дискретної змінної або закон розподілу щільності ймовірності неперервної випадкової змінної

Корисно пам'ятати

Ймовірність, що $x < a$	$F(a)$
Ймовірність, що $x = a$	$p(a)$
Ймовірність, що $x > a$	$1 - F(a)$
Ймовірність, що $a < x < b$	Ймовірність, що $x < a$ $F(a)$ Ймовірність, що $x < b$ $F(b)$ Попадання в інтервал від a до b - ймовірність другого мінус ймовірність першого $F(b) - F(a)$

Деякі закони/ функції розподілу

Дискретні

Рівномірний розподіл

- Орел-решка
- кыдок однієї кістки

Розподіл Бернуллі (біноміальне)

- Серія незалежних випробувань, в кожному з яких подія A може з'явитись з однаковою ймовірністю p
- Кость кидають 5 разів, яка ймовірність двох шісток?

Рівномірний розподіл Пуассону)

- випадкова змінна – число подій, які відбулися за фіксований час, за умов, що ці події виникають з певною фіксованою інтенсивністю
- Кількість дзвінків до реєстратури

Геометричний (до першого успіху)

- Гіпергеометричний (кількість успіхів без повернення)
- Від'ємне біноміальне (кількість невдач)

Безперервні

Параметри, що характеризують випадкові змінні

Важливо!

- Вважається, що всі дані, з якими ми працюємо - частини генеральної сукупності (вибірки з генеральної сукупності)
- Генеральна сукупність має закономірні властивості
- По вибіркам ми намагаємося уявити ці загальні властивості генеральної сукупності, тобто робимо розрахунок:
- вибіркових характеристик
- оцінок параметрів

Мода

- Найчастіше значення випадкової змінної
- Мода випадкового розподілу діагнозів по поліклініці за день
- Мода суми значень, викинутих на двох кістках-кубиках
- Мода орел чи решка

Середнє арифметичне(mean)

- Вибіркова оцінка математичного очікування
- Центральний момент вибірки першого порядку
- Найбільш часта характеристика вибірки
- $m_x = \frac{1}{N}\sum_{i=1}^{i=N} x_i$

- Недолік: при асиметричних даних не завжди потрапляє в дійсний центр даних

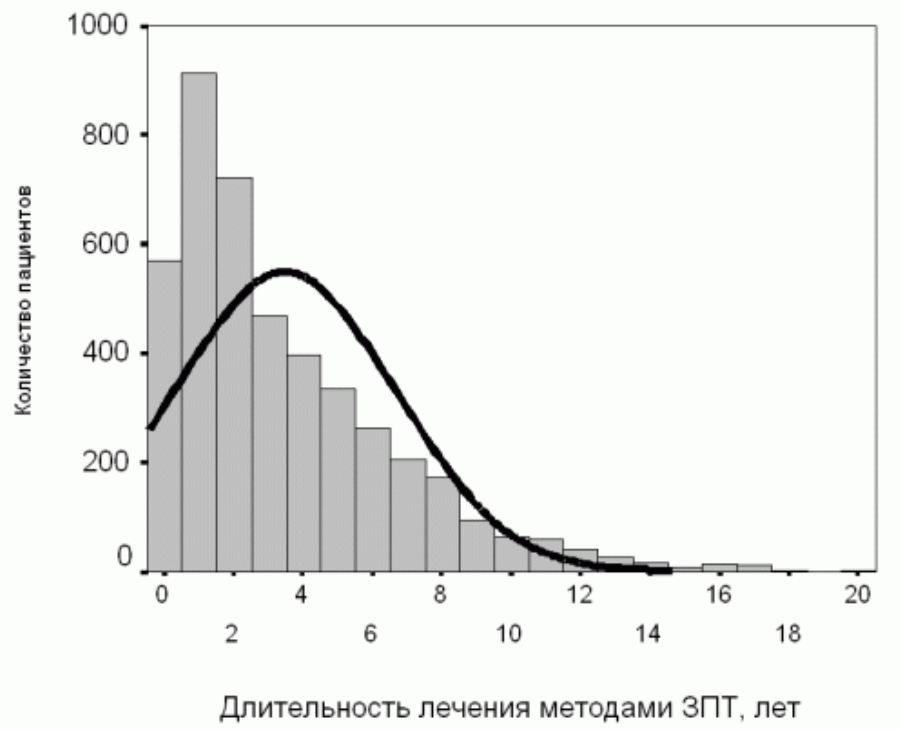
Медіана, квартилі

- Нехай наші дані розташовані по зростанню побудованого варіаційного ряду
- По осі Y - цілі числа від 0 до N (або якщо їх поділити на N - о відсотки від 0 до 100%)
- Медіана і т.п. - це просто характерні мітки на осі Y

Медіана, квартилі

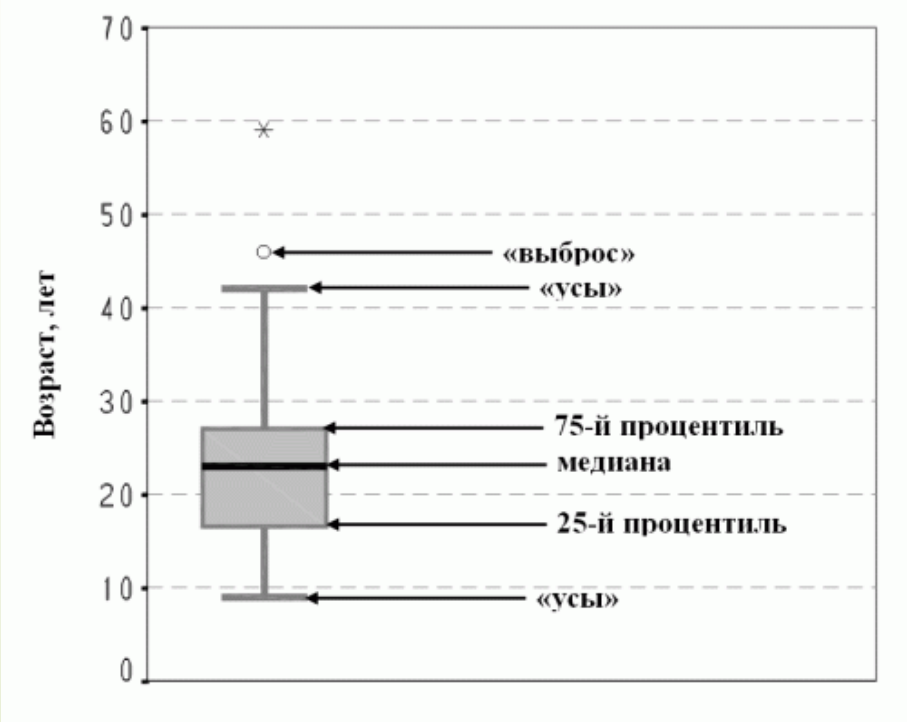
- Медіана - такий x, якому на осі Y відповідає рівно 50%
- Квартилі - 25%, 50%, 75%
- 25% - нижній квартиль
- 50% - медіана - середній квартиль
- 75% - верхній квартиль

Медіана та квартилі



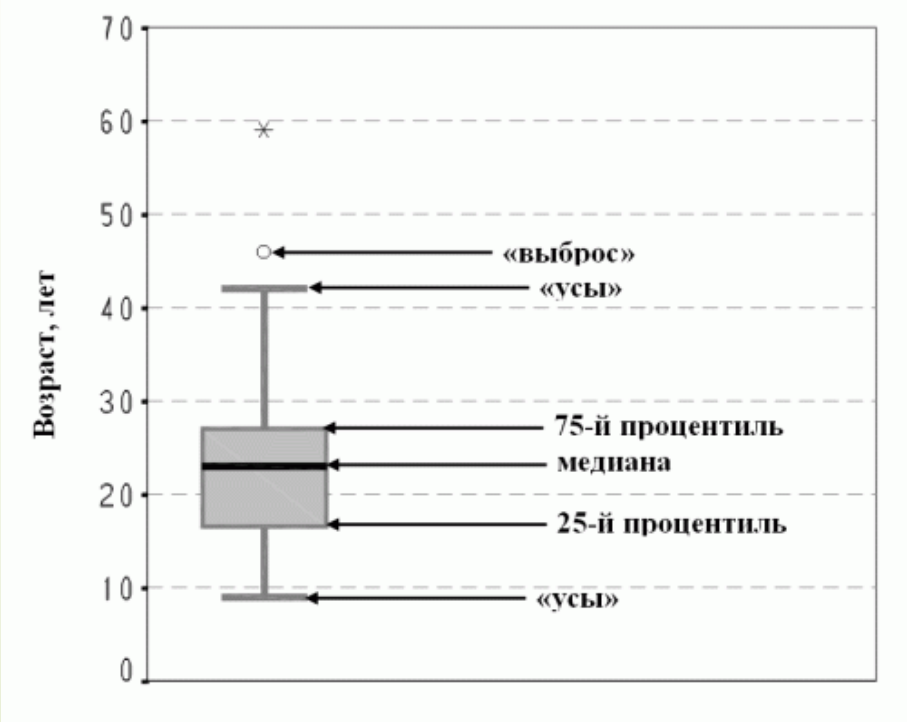
- Середній термін спостереження (арифметично - 3,5 року)
- Медіана - 2,5 року
- Що «більш правильно»?

Boxplot☺



- Ящичок зафарбований сірим кольором.
- Його нижня межа є 25 процентіль, верхня - 75 процентіль.
- Горизонтальна чорна риса, яка перетинає скриньку - це медіана.
- Як бачимо, медіана ділить скриньку на дві нерівні частини - значить в розподіл, відображене на малюнку, носить неправильний характер.

Boxplot ☺



- Від boxplot відходять «вуса».
- У прикладі на малюнку нижній «вус» відображає інтервал, в якому перебувають 25% найнижчих значень - від 9 до 17.
- Слід звернути увагу, що над верхнім «вусом» є дві точки - викиди.
- Тому верхній «вус» відображає інтервал, в якому перебувають 25% мінус викиди, які складають 2,8%.
- Таким чином, якщо викидів немає, то «вус» відображає інтервал, в якому перебувають 25% всіх спостережень.
- Якщо ж викиди є, то «вус» відображає інтервал, в якому знаходяться значення від квартили до величини, яка менше, ніж півтори довжини шухлядки

Дисперсія (розкид) та інші характеристики

Розкид (Variation)

- Дисперсія - міра розкиду даних щодо середнього
- Оцінка за вибіркою (середній квадрат відхилення):
- $S_x^2 = \frac{1}{N-1}\sum_{i=1}^{i=N}(x_i - m_x)^2$
- Примітки:
- Просто сума відхилень вліво-вправо при підсумовуванні дало б 0 або щось близьке
- Можна складати модулі, але це незручно в подальших розрахунках
- Сума квадратів - зручніше ніж модулі

Стандартне відхилення(Std Dev)

- $StdDev = \sqrt{Variance}$ та навпаки
- Одиниця виміру - та ж, що у випадкової величини
- Якщо розподіл схоже на нормальне, то 99% даних

Розмах (Range)

- $Range = x_{max} - x_{min}$
- Є випадки коли використовується☺

Межквартільний інтервал

- Відстань між верхнім і нижнім квартилем, охоплює 50% випадкових величин у вибірці (див. Вище аналіз асиметричною вибірки)

Асиметрія (skewness)

Важкий лівий хвіст розподілу – skew>0;
Важкий правий хвіст – skew<0;
Skew=0 – щільність розподілу **симетрична**

- Центральний момент третього порядку
- $Sk = \frac{1}{N}\sum(x_i - m_x)^3$
- Коефіцієнт асиметрії - то ж, тільки поділене на куб стандартного відхилення
- $K_{Sk} = 1/\sigma^3 \frac{1}{N}\sum(x_i - m_x)^3$
- Коефіцієнт асиметрії **позитивний, якщо правий хвіст розподілу довше лівого, і негативний в іншому випадку**.
- Якщо **розподіл симетрично** щодо математичного очікування, то його **коефіцієнт асиметрії дорівнює нулю**.