

## 1. Checking assumptions for applying the t-test

You need to answer the following question and provide a method for testing the assumption:

Is it necessary for the *original data* to be normally distributed in order to apply the t-test?

Different sources provide different interpretations. Some emphasize that the *sampling distribution of the test statistic* should be approximately normal, while others state that the *original data itself* must follow a normal distribution.

The result is expected in the form of a clear **answer to the question** with a brief explanation, as well as **working code** that implements a check for whether the assumption holds for a given dataset.

## 2. Analysis of the user behaviour

You need to analyze the logs attached to the task and determine how user behavior differs between the two samples (each log file represents one sample).

You should also interpret the identified difference.

The result is expected in the form of a **working code (script or notebook)** performing the analysis, along with the **conclusion** based on the results.

Log format (`users_log_raw_a(b)10000.txt`):

The logs include **Session metadata**, **Query**, and **Click actions**.

- `TypeOfRecord` — type of the log entry: query (Q), click (C), session metadata (M).

Session metadata (TypeOfRecord = M):

Fields: `SessionID \t TypeOfRecord \t Day \t USERID`

- `SessionID`: unique session identifier
- `Day`: day number when the session occurred
- `USERID`: unique user identifier

### Query actions (TypeOfRecord = Q):

Fields: SessionID \t TimePassed \t TypeOfRecord \t SERPID \t QueryID \t ListOfTerms \t ListOfURLsAndDomains

- **SessionID**: unique session identifier
- **TimePassed**: time units passed since the beginning of the session (the unit duration in milliseconds is unspecified)
- **SERPID**: unique search results page identifier
- **QueryID**: unique query identifier
- **ListOfTerms**: list of search terms in the query
- **ListOfURLsAndDomains**: ordered list of URLs with domains shown on the search results page
  - Format: (URLID, DomainID) — unique URL identifier, unique domain identifier

### Clicks (TypeOfRecord = C):

Fields: SessionID \t TimePassed \t TypeOfRecord \t SERPID \t URLID

- **SessionID**: unique session identifier
- **TimePassed**: time units passed since the beginning of the session (the unit duration in milliseconds is unspecified)
- **SERPID**: unique search results page identifier
- **URLID**: unique URL identifier that was clicked

**Clicks are attributed to the QueryID that precedes the clicks in the log.**