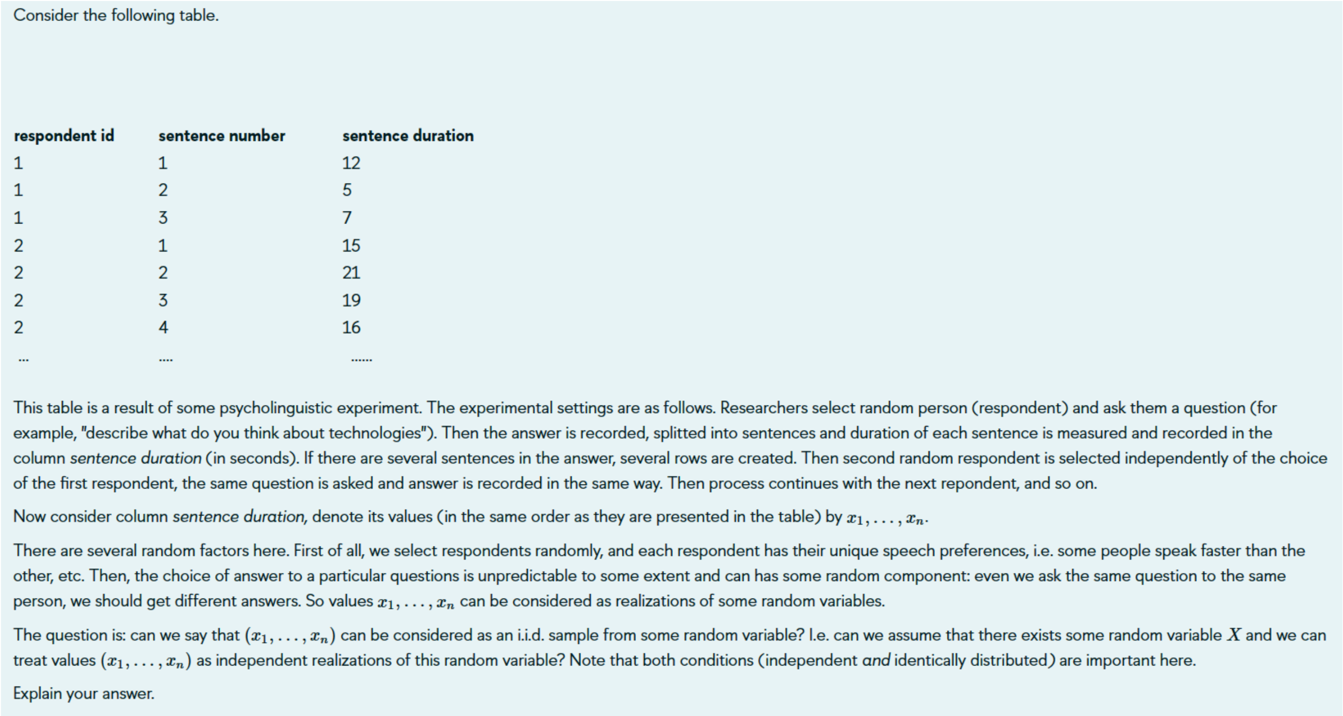
Question 2



Consider the following table.

respondent id	sentence number	sentence duration
1	1	12
1	2	5
1	3	7
2	1	15
2	2	21
2	3	19
2	4	16
...	....	......

This table is a result of some psycholinguistic experiment. The experimental settings are as follows. Researchers select random person (respondent) and ask them a question (for example, "describe what do you think about technologies"). Then the answer is recorded, splitted into sentences and duration of each sentence is measured and recorded in the column *sentence duration* (in seconds). If there are several sentences in the answer, several rows are created. Then second random respondent is selected independently of the choice of the first respondent, the same question is asked and answer is recorded in the same way. Then process continues with the next repondent, and so on.

Now consider column *sentence duration,* denote its values (in the same order as they are presented in the table) by $x_1, \ldots, x_n$.

There are several random factors here. First of all, we select respondents randomly, and each respondent has their unique speech preferences, i.e. some people speak faster than the other, etc. Then, the choice of answer to a particular questions is unpredictable to some extent and can has some random component: even we ask the same question to the same person, we should get different answers. So values $x_1, \ldots, x_n$ can be considered as realizations of some random variables.

The question is: can we say that $(x_1, \ldots, x_n)$ can be considered as an i.i.d. sample from some random variable? I.e. can we assume that there exists some random variable X and we can treat values $(x_1, \ldots, x_n)$ as independent realizations of this random variable? Note that both conditions (independent *and* identically distributed) are important here.

Explain your answer.

We cannot consider $(x_1, ..., x_n)$ as an i.i.d. sample from some random variable.

1) Non-independence:
The durations of sentences spoken by the same person may not be independent from each other. For example, if a person tends to speak faster or slower than average, this may affect the duration of all of his/her sentences, not just one.

An example of the dependency between sentence durations within a respondent's answer can be illustrated through the following scenario. Let's consider event A, where the duration of the first sentence in a respondent's answer is longer than some average duration, and event B, where the duration of the second sentence is longer than some average duration. If we assume that these events are independent, then the probability of event B given event A should be equal to the probability of event B, or $P(B|A) = P(B)$. However, in reality, the occurrence of event A provides information about the speaker's tendencies in sentence duration, which can impact the probability of event B. For example, if a person tends to speak more slowly in general, then the occurrence of event A (a longer than average first sentence) may increase the probability of event B (a longer than average second sentence), as the person may continue to speak at a slower pace throughout the response.

There may be other patterns related to a respondent's speaking pace, level of fluency with the language, or familiarity with the topic being discussed, that would allow to predict next sentences duration basing on the first ones. Therefore, it is important to account for the dependency between sentence durations within a respondent's answer, rather than assuming independence, when analyzing and modeling data from psycholinguistic experiments.

2) Non-identical distribution:
The sentence durations may not be identically distributed because the number of sentences in a response can be arbitrary, which may affect the distribution of sentence durations in the sample.

For example, more talkative people may have a stronger impact on the distribution because they contribute more values to the sample. Suppose a talkative respondent on average produces a response with five long sentences, while a less talkative respondent produces an average response with only two sentences of moderate length. Then, the distribution of sentence durations in the sample may be skewed towards longer durations due to the contribution of talkative respondents. Therefore, the assumption of identically distributed sentence durations may not hold in this case.