1

- the participants' times are, like you suggested, drawn from a non-normal distribution.
- however, note that when you are running the t-test you are **not** comparing *individual times* from either set (comment/no comment) to individual times from the other set.
- in your t-test,, the test statistic is *mean of the observation times* of one set to the mean of the observation times of the other set. meaning, you are you are comparing the means, so the only thing that matters is whether the means are normally distributed.
- it is an underlying assumption of the t-test that the test statistic is normally distributed. is that so in your case? i would say that it is likely (especially if your number of respondents is high).
- you **should** verify that. how can you check that the mean is normally distributed, when you only have the one mean? i suggest you take a large number of random samples from your set of observations (take care to have every value equally likely to get sampled and sample with replacement). for each sample, compute its mean. plot the histogram of these means and check to see whether it is bell shaped, symmetric, centered on the mean of the full set.
- when you do that sampling, optimally (and quite likely), each set for each question has a bell shaped histogram of means. but the two sets you are comparing might have different variances. if so, then that violates one of the assumptions of the t-test.
- you may be able to determine that by eye, but if not there are statistical tests to check whether variances of two samples are comparable. (in that case we can talk more on those later).
- if you are convinced that the variances differ between the sets (comment/no comment) then the (plain) t-test is not appropriate and might give misleading results.
- in this case, there are variations of the t-test, namely welch's test.
- are you running these tests using python `scipy.stats.ttest_ind(a, b, axis=0, equal_var=False)` ? or R `t.test(no_comment_times, with_comment_times, alternative="two.sided", var.equal=FALSE)` ? or something else? # 2
- i am not sure i know the "standard Median, Q1 and Q2 methodology" but i know sometimes outliers are identified using a method due to tukey, where outliers are defined as points more than 1.5 times the interquartile range away from the first and third quartiles. ie one drops the points that lieeither
  - below `q1 – 1.5×iqr`
  - or above `q3 + 1.5 iqr`

- though commonly applied, i am not aware of any principled reasoning that convinces me the that this method is reliable. to drop outliers you have to have a model for how they come about. in you case, i expect you would have a handful of time measurementst that are super long and these are likely due to the participants getting distracted or bored and doing something else for a while. in that case i would expect it to be reasonable to use something like tukey's method. # 3
- i agree with your colleague to try the chi-squared test. in it, your null hypothesis is that the success rate is independent of the presence of the comment, ie that both ratios were drawn from the same underlying distribution. the chi-squared test computes a contingency table and evaluates the four numbers:
  - the expected number of people that would get the **correct** answer *without the comment*,
  - the expected number of people that would get the **incorrect** answer *without the comment*,
  - the expected number of people that would get the **correct** answer *with the comment*,

    - the expected number of people that would get the **incorrect** answer *with the comment*,

- these expected numbers are compared with the actual counts found and the total deviation from expectation is your test statistic.