greatlearning
Power Ahead

SMDM Week 3 FAQs

Q1. Why is it that we divide p value by 2 when we do a 1 sample test?

Ans. *Python outputs a two-sided/ two tailed p-value by default. Whenever we are doing a one-tailed test, we are interested in only one side of the p-value, left tail or the right tail. Because of the limitations of Python, it is a standard practice to divide p-value by 2 for one tailed test. However, here, t-statistics value needs to be checked too. If t-statistics value is less than 0 and you are looking for a right tailed test, then you would need to also subtract the (p-value/2) from 1.*

Q2. Please help identify between one tailed and two tailed tests.

Ans. *Let's say a problem states that the average response from the population is 5 and a research scientist claims that the average response is different from the population average response.Here, the word "different" means that it can be either greater or lesser than the average response that suggests it is a two-tailed.If the problem, for example, states that the average response from the population is let's say 5 and a research scientist claims that the average response is less than the population average response.Here, the word "less" means that it is less than the average response that suggests it is a one-tailed & left tailed test.If the problem, for example, states that the average response from the population is let's say 5 and a research scientist claims that the average response is greater than the population average response.Here, the word "greater" means that it is more than the average response that suggests it is a one-tailed & right tailed test.*

Q3. Please help write the steps for a paired t-test in python.

Ans. *The steps are listed below:*
1. Formulate the Null and Alternate Hypothesis
2. Identify the test to be performed, which type (one-tailed or two-tailed), and level of confidence.
3. Calculate the critical value, degrees of freedom.
4. Calculate the test statistic.
5. Infer the results.
Please refer to these links for more info:
https://pythonfordatascienceorg.wordpress.com/paired-samples-t-test-python/https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

Q4. Please help understand p-value it the probability of an event in the sample taken, given the H0 is true?

Ans. *p-values denotes the probability value, i.e., the chances that a particular event is going to happen. Like the chances of Head occurring on a coin toss, like the probability that a particular movie crossing 100 crores in revenue, like a particular video going viral, like the likelihood of rain today. Each of these events have a chance of happening, and this chance is called probability or the p-value. If we put it in context of Hypothesis testing, then p-value means the chances of the sample value being the same as the observed (population value). Suppose there is a H0 that the share price of Yes Bank will be 50 today. Our approach would be to take a sample of a few days and observe the share prices of this particular stock. Some days it will be more than 50, some days it will be less than 50 but each day will be a different value. So ideally, we would take an average of say the last 100 days of the stock price of Yes Bank and compare it with the given value (50). Let us say the average comes out to be 48. Please note that this value is derived from the latest 100 days of the prices of the stock. This may or may not be true for the entire duration of the stock (i.e. since the day Yes Bank was listed on the stock exchange until today = population and 100 days =sample). Now we find the chances of our sample mean (48) being equal to the claim (50). The chances of the mean being 50 is represented by p-value. If this p-value (probability that the price is 50) is less than 0.05 (default significance level) then we reject H0. Let's say the p-value comes out to be 0.04 (means that there is*

4% chance that the share price of Yes Bank today will be 50), we will reject the H0 that the share price will be 50 today.

Q5. What is the purpose of Wilcoxon test?

Ans. *The Wilcoxon signed-rank test tests the null hypothesis that two related paired samples come from the same distribution. In particular, it tests whether the distribution of the differences x - y is symmetric about zero. It is a non-parametric version of the paired T-test. For more info: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html*

Q6. What does fat tail in a distribution mean? For example, it is said that t-distribution has a fat tail with small sample size. Is it different from heavy tailed distribution?

Ans. *A distribution with heavier/Fatter tails is a distribution where the probability of being far from the mean drops slowly. In terms of Nasdaq -By definition, a fat tail is a probability distribution which predicts movements of three or more standard deviations more frequently than a normal distribution.*

Reference: https://www.nasdaq.com/articles/fat-tail-risk-what-it-means-and-why-you-should-be-aware-it-2015-11-02

The "heavy tail" and "fat tail" mean the same thing, and is common in trading and other areas of finance. For more info Tail Risk (in terms of trading): https://www.investopedia.com/terms/t/tailrisk.asp

Q7. Total Fat (% Daily Value) what is the meaning of this column. Does it show the % present in the diet or should be present in the diet. i.e. ideal state.

Ans., The *percentage daily value (%DV) shows how much a nutrient in a serving of food contributes to a total daily diet. The %DV can help you determine if a serving of food is high or low in a nutrient.*

Q8. What is p-value in a very simple layman terms & why we need to compare with Alpha and not Beta

Ans. *The p-value is the probability that the null hypothesis is true. p-values tell us whether an observation is a result of a change that was made or is a result of random occurrences. For example, consider that a class has n number of students and they have been asked to attempt a quiz twice. 1 quiz is to be attempted before them having attended a session on the topic of the quiz and the other quiz is to be attempted after having attended the session on the topic of the quiz. After taking up both the quizzes, the teacher is trying to find out whether there is any significant difference between the scores of the students in both the quizzes. To verify this, the teacher selects a sample of students from all the students and compares their scores in both the quizzes. Here, an important thing to note is that teacher is trying to claim that there might be a difference in the scores of the 2 quizzes. So, we can assume the current fact to be that there is no significant difference between the scores of the 2 quizzes. So, here the hypothesis will be as follows:*

H0: There is no significant difference between the marks obtained by a student in both the quizzes

Ha: There is a significant difference between the marks obtained by a student in both the quizzes

Say, we get a p-value = 0.2. That means, the probability that our null hypothesis is true is 0.2 i.e., we are 20 percent sure that there is no significant difference between the marks obtained by a student in both the quizzes. Now, setting up a cut off value(alpha) is upon us. So, how aggressive do we want to be in terms of rejecting the fact that there is no significant difference in the marks. If we think having less that just having 30 percent surety for null hypothesis can be alarming, then we can set alpha = 0.3. Generally, alpha = 0.05 is just a standard convention used and this can differ for industry to industry.

Q9. If the sample size is not same in paired data (i.e.) in one sample it has No values- how we will do a T Test. As per the 'Step 3- Identify Statistics' in the T Test (two- sample).ipynb attached in the Week 3 of Statistical (Hypothesis Testing), to do a paired data, the sample size should be equal.

Ans. *A paired t-test is applied in a case where there is some sort of experiment performed and there are 2 observations for each case. One observation will represent the value that occurs before the experiment and the other observation will be after the experiment. So, there will always be before and after values available. Therefore, for a paired data, the sample size will be equal.*

Q10. While trying to use t statistics, we assume that the variances are equal. Why are we making such an assumption? Please explain. Moreover, when we say variances are equal do we mean variance between 2 samples are equal or between populations.

Ans., *If the variances are unequal then the test of hypothesis is more susceptible to errors. For the second part, we are only considering the equality or the inequality of the population variances.*