When to use T-test, Chi square test, ANOVA test

**More focus on**: This is a quick review note, just took your 10 mins. and through this note, we all not only be **able to understand when to use which test but also how to set the null hypothesis and alternative hypothesis.**

We all often get confuse when to use T-test, Chi square test, ANOVA test while we are given any dataset and have to test whether null hypothesis is true or alternative hypothesis is true.

***Note*: How to know by looking the data and by considering the requirements of the data which test is the best or appropriate for the effective result.**

Let us consider a simple and random dataset for better understanding;

Pay attention: It is just for understanding and basic statistics theory problems can be solved.

Let’s get started:

THEORITICAL UNDERSTANDING OF THE DATASET

In this dataset, I will be considering two columns as categorical data and other two are numerical data

**special observation:** This dataset consists of **four feature or columns as Age Group, Gender, Height and Weight.**

Let me tell one more important thing about this data:

further segregation of feature or columns (data science machine learning term)

Here,

**Age Group and Gender are categorical data and both are Nominal data**

**Height(m) and Weight(kg) are numerical continuous data.**

This is a sample dataset

I will be providing test situation by considering questions based on the data (which might be asked in real time situations as well), for better understanding of the usage of each test.

Dataset

|  |  |  |  |
| --- | --- | --- | --- |
| GENDER | AGE GROUP | HEIGHT (m) | WEIGHT(KG) |
| MALE | YOUNG ADULTHOOD | 1.4 | 40 |
| FEMALE | YOUNG ADULTHOOD | 1.3 | 45 |
| MALE | OLD ADULTHOOD | 1.4 | 80 |
| MALE | ADOLESCENCE | 1.2 | 30 |
| FEMALE | YOUNG ADULTHOOD | 1.5 | 55 |
| FEMALE | OLD ADULTHOOD | 1.3 | 60 |

**Question1**: Now, consider one feature column in AGE GROUP, that what is the difference between the proportion of Male and Female.

So why this let us set null hypothesis and alternative hypothesis

|  |  |
| --- | --- |
| Feature or column | GENDER (One Categorical data) |
| Null hypothesis | There is no difference in proportion of male and female |
| Alternative hypothesis | There is a difference in proportion of male and female proportion |
| Test | One sample proportion test |

Assumption: we always consider null hypothesis to be true and then see what likelihood or probability of alternative hypothesis to be true.

**Curious question:**

When we can set alternative hypothesis that there is a difference than what is the need of null hypothesis while testing or vice versa.

**Reasons:**

1. **This is a sample data obvious we need null hypothesis as the case of errors that may arise.**
2. **In test, if we have to proof something than for setting it is true, we need something to disproof, to claim our result.**

Why have I arisen this question, because by drawing bar graph we can conclude that there is a difference (alternative hypothesis), still we are setting null hypothesis for testing. So, I hope now the reason is clear.

**How to draw conclusion:**

Just test the two hypothesis and see whether the value of test lies in the rejection region or do not accept region.

**If we use: P- value than,**

1. **P-value less than 0.05 (significance value or alpha) than reject null hypothesis (the value lies in the rejection region in the tail part)**
2. **P-value greater than 0.05 than do not reject null hypothesis or accept alternative hypothesis (value lies in the center part of the normal distribution)**

**Following above question1, the rest question can be answered by using same assumption, knowledge and how to draw conclusion.**

**Question2**: considering two categorical features or columns and what to test that is there any difference of Gender (male or female) based on Age Groups?

|  |  |
| --- | --- |
| Feature or column | GENDER and AGE GROUP (Two Categorical data) |
| Null hypothesis | There is no difference in proportion of male and female based on Age Group |
| Alternative hypothesis | There is a difference in proportion of male and female proportion based on Age Group |
| Test | CHI SQUARE TEST |

**Question3**: considering weight and we need to test whether the mean from the previous data is different or not?

|  |  |
| --- | --- |
| Feature or column | WEIGHT (Numerical data) |
| Null hypothesis | There is no difference in the mean value  Mean = 51.67 |
| Alternative hypothesis | There is a difference in the mean value  Mean 51.67 |
| Test | T-TEST |

Mean of the above data is 51.67

Yes, you are thinking correctly that in the case of two numerical data test we can use T-test

**But, have to see conditions whether to apply Z-test or T-test:**

|  |  |
| --- | --- |
| For t-test   1. Sample size should be less than 30 2. Population variance is unknown | For Z-test   1. Sample size should greater or equal to 30. 2. Population variance is know |

**Question4**: if we have to test two numerical data whether there is any relationship between each other?

|  |  |
| --- | --- |
| Feature or column | WEIGHT and HEIGHT (Two Numerical data) |
| Null hypothesis | There is no relationship between the two data |
| Alternative hypothesis | There is a relationship between the two data |
| Test | CORRELATION  Using T-Test |

Now, how to see conclusion, (I have already made notes on correlation in detail)

Correlation value lies between -1 to 1 and if the result or test value comes as 0 than that means there is no correlation and we will do not reject null hypothesis.

**Yes, correct value can be determined by using T-test in our example.**

**Question5**: now, considering one numerical data and two categorical data if we have to come to the conclusion than use:

|  |  |
| --- | --- |
| Feature or column | WEIGHT (Numerical data) and AGE GROUP, GENDER (Categorical data) |
| Null hypothesis | There is no relationship |
| Alternative hypothesis | There is a relationship |
| Test | ANOVA TEST |

**Wrong assumption**: that we can apply this test just by knowing that we have one numerical and two categorical data and you will everywhere use ANOVA test.

**Why ANOVA test?**

**Here, gender has more than 2 categories like ADOLESCENCE, YOUNG AND OLD ADULTHOOD, there we use ANOVA test.**

When the case having 2 or less than 2 than use T-test (in our example).