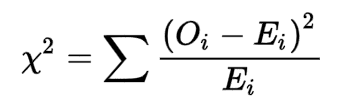
**QUIZ-1**

1. **Measures of Central Tendency.**
   1. **Mean example vs median example**
      1. **For using mean as a measure**:
         1. An example dataset (distribution) where mean is the preferred measure would be distribution of temperatures in a city or a town. For example let us consider a dataset for a town as [40, 45, 50,60,70,80,85,85,75,65,55,45] where it contains 12 entries for each month of the year. The average temperature(mean) would be sum(all values) / 12 =~ 65 which is average monthly temperature. So while analyzing the data for getting average monthly temperature, mean would be the preferred choice as it provides the better representation for average temperature in a year..
      2. **For using median as a measure**:
         1. An example dataset where we would prefer median as a measure as supposed to mean would be data regarding home/houses prices in a locality. Consider the dataset containing values for various houses in a city ranging from cheap apartments to high end luxurious villas. For example,Here, mean would be inefficient as the luxurious apartments are the outliers (majority of people come in lower or middle class)and by including outliers would heavily impact the outcome. Here, median works better as it provides the middle element and is able to deal with heavy outliers
   2. **Mode**:
      1. For some datasets, we can only use mode is when the dataset doesn't lend itself to meaningful mathematical operations and has multiple values/categories. An example of such a dataset would be Preferred Clothing Brand. In this case, the dataset contains non numerical values where we cannot perform mathematical operations and thus mean and median would not work.
2. **Correlation:**
   1. The value of the Pearson correlation coefficient between variable X and Y is -0.95. Are this variables correlatied? Explain.
      1. Since the value of Pearson correlation coefficient between variable X and Y is -0.95, it suggests that there is strong negative correlation between both the variable. Strong negative correlation means that both variables move in opposite direction meaning simply if X increase, Y decreases.
      2. An example would be relation between coffee consumption and productivity of a person.Here, the confounding variable is sleep quality where Coffee drinkers who experience poor sleep might consume more coffee to compensate, making it essential to consider sleep quality's influence.
3. **Chi-square correlation test:**
   1. Formulate the null hypothesis:
      1. The null hypothesis states that there is no correlation between the tow variables here they are age and ability to swim. Therefore, the null hypothesis here would be there is no correlation between age and ability to swim.
   2. Find the chi-square value

| OBSERVED | Can Swim | Can't Swim | Total |
| --- | --- | --- | --- |
| Age <= 20 | 58 | 36 | 94 |
| 20 < Age <= 40 | 68 | 38 | 106 |
| 40 < Age <= 60 | 50 | 45 | 95 |
| Age > 60 | 22 | 20 | 42 |
| Total | 198 | 139 | 337 |

| EXPECTED | Can Swim | Can't Swim | Total |
| --- | --- | --- | --- |
| Age <= 20 | (94\*198)/337 =55.2 | (94\*139)/337=  38.7 | 94 |
| 20 < Age <= 40 | (106\*198)/337 = 62.2 | (106\*139)/337 =43.7 | 106 |
| 40 < Age <= 60 | (95\*198)/337 = 55.81 | (95\*139)/337 = 39.1 | 95 |
| Age > 60 | (42\*198)/337 = 24.67 | (42\*139)/337 = 17.32 | 42 |
| Total | 198 | 139 | 337 |

* + 1. 

= (58-55.2)^2/55.2 + ………so on

= 3.82

* 1. Assume the significance level to be 0.025. Should you accept or reject the null hypothesis? Justify.
     1. The degree of freedom when calculated using the formula (r-1)\*(c-1), we get 3.Since the value we found above is lower than the critical value which is 9.3, therefore we should accept the null hypothesis.
  2. According to the test results, are age of the participants and ability to swim correlated? Justify.
     1. No, they are no related as we proved earlier as the chi value is less than the critical value and we have accepted the null hypothesis.