Problem Statement - Answer the following questions to the best of your knowledge including the concepts taught to you in the level.

1. On what basis we choose data scaling method (Normalization/Standardization)?

2. If the VIF is 2 then what is value of correlation coefficient (r^2)

3. How do you interpret chi-square result?

4. Why do we choose boxplot method than other for outlier detection and removal?

5. How do we choose best method to impute missing value for a data?

1. On what basis we choose data scaling method (Normalization/Standardization)?

Solution:

Feature Scaling methods are used to to keep all the features in a dataset within a range so that none of the variable dominate in predictions of the model.

Following Techniques are used for feature Scaling:

1. Normalization:

It focuses on rescaling the values of features in the range from 0 to1 without distorting differences between them. Normalization uses the minimum and maximum values of the feature to accomplish rescaling in range of 0 to 1 . Thus using normalization we can bring dataset to a common scale.

Xnormalized = (Xi - Xmin) / (Xmaximum - Xminimum)

1. Standardization:

Standardizing the features around the center and 0 with a standard deviation of 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating abais.

Xstnadardizes= (XI - XMEAN) / XSD

We should consider following points while selecting scaling method:

1. The data distribution :

Data distribution is very important factor to be considered in selecting scaling method.

If the data is normally distributed (Bell Curve) i.eif data lies equally to both the sides of mean at given multiples of SD then data is normally distributed.

If data is normally distributed then we should go for Standardization and not Normalization. Since Standardization will give us location ofevery data-point in multiple of SD from mean.

1. Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors and artificial neural networks.
2. Standardization assumes that data is normally distributed so it must be used when the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression, and linear discriminant analysis.

2. If the VIF is 2 then what is value of correlation coefficient (r^2)

Answer:

Pearson’s correlation coefficient (R) is an indicator which shows linear relationship between two variables.

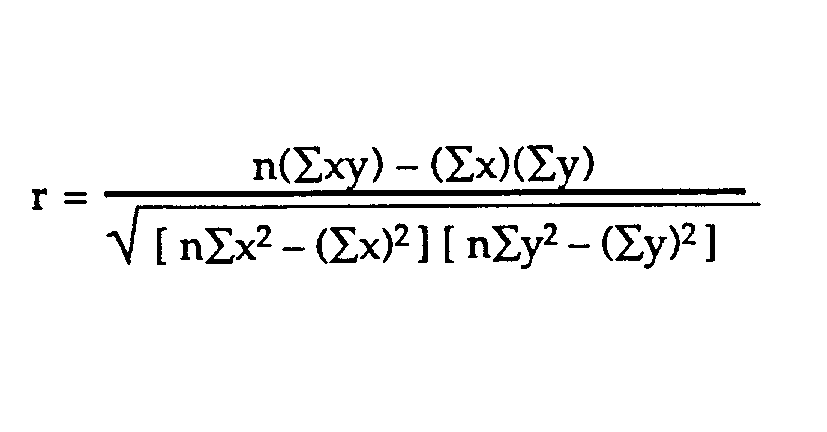
Range of R is -1 to 1

-1 denotes extreme negative correlation

0 denotes no correlation

1 denotes extreme positive correlation

It is calculated as follows:



VIF (variance Inflation Factor ):

Variance Inflation Factor (VIF) is a common simple stat used to quantify multicollinearity in regressions. It is calculated for each covariate in a regression, with higher values meaning that the covariate is more colinear with the other covariates. It [technically](https://en.wikipedia.org/wiki/Variance_inflation_factor)measures “how much the variance of an estimated regression coefficient is increased because of collinearity.” The equation is:

VIF = 1/1-R2

If ,

VIF = 2

Then,

2 = 1/1-R2

2(1-R2) = 1

2 – 2R2 = 1

2R2 = 1

**R2  =1/2**

3. How do you interpret chi-square result?

**Answer:**

Following Steps must be followed while doing Chi-Square test.

## State the Hypotheses

Suppose that Variable A has r levels, and Variable B has c levels. The [null hypothesis](https://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

Ho: Variable A and Variable B are independent.

Ha: Variable A and Variable B are not independent.

The [alternative hypothesis](https://stattrek.com/Help/Glossary.aspx?Target=Alternative%20hypothesis) is that knowing the level of Variable A can help you predict the level of Variable B.

## Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

* Significance level. Often, researchers choose [significance levels](https://stattrek.com/Help/Glossary.aspx?Target=Significance%20level) equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
* The significance level is the probability of rejecting the null hypothesis when the null hypothesis is true.: It’s the probability of making a wrong decision.

## Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

* **Degrees of freedom.** The [degrees of freedom](https://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom) (DF) is equal to:

DF = (r - 1) \* (c - 1)

where r is the number of levels for one catagorical variable, and c is the number of levels for the other categorical variable.

* **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute r \* c expected frequencies, according to the following formula.

Er,c = (nr \* nc) / n

where Er,c is the expected frequency count for level r of Variable A and level c of Variable B, nr is the total number of sample observations at level r of Variable A, nc is the total number of sample observations at level c of Variable B, and n is the total sample size.

* **Test statistic.** The test statistic is a chi-square random variable (Χ2) defined by the following equation.

C2 = Σ [ (Or,c - Er,c)2 / Er,c ]

where Or,c is the observed frequency count at level r of Variable A and level c of Variable B, and Er,c is the expected frequency count at level r of Variable A and level c of Variable B.

* **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. i.e  the probability of obtaining a result as extreme as, or more extreme than, the result actually obtained when the null hypothesis is true.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the [significance level](https://stattrek.com/Help/Glossary.aspx?Target=Significance%20level), and rejecting the null hypothesis when the P-value is less than the significance level.

Example:

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Prefrences** | | | **Row total** |
| **BJP** | **Congress** | **Other** |
| **Male** | 200 | 150 | 50 | 400 |
| **Female** | 250 | 300 | 50 | 600 |
| **Column total** | 450 | 450 | 100 | 1000 |

We use 0.05 level of significance.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the [null hypothesis](https://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) and an alternative hypothesis.

Ho: Gender and voting preferences are independent.

Ha: Gender and voting preferences are not independent.

* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [chi-square test for independence](https://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20test%20for%20independence).
* **Analyze sample data**. Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the [degrees of freedom](https://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom), we determine the [P-value](https://stattrek.com/Help/Glossary.aspx?Target=P-value).

DF = (r - 1) \* (c - 1) = (2 - 1) \* (3 - 1) = 2

Er,c = (nr \* nc) / n  
E1,1 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,2 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,3 = (400 \* 100) / 1000 = 40000/1000 = 40  
E2,1 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,2 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,3 = (600 \* 100) / 1000 = 60000/1000 = 60

Χ2 = Σ [ (Or,c - Er,c)2 / Er,c ]  
Χ2 = (200 - 180)2/180 + (150 - 180)2/180 + (50 - 40)2/40  
    + (250 - 270)2/270 + (300 - 270)2/270 + (50 - 60)2/60  
Χ2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60  
Χ2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2

where DF is the degrees of freedom, r is the number of levels of gender, c is the number of levels of the voting preference, nr is the number of observations from level r of gender, nc is the number of observations from level c of voting preference, n is the number of observations in the sample, Er,c is the expected frequency count when gender is level r and voting preference is level c, and Or,c is the observed frequency count when gender is level r voting preference is level c.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

 P(Χ2 > 16.2) = 0.0003.

* **Interpret results**. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

4. Why do we choose boxplot method than other for outlier detection and removal?