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.

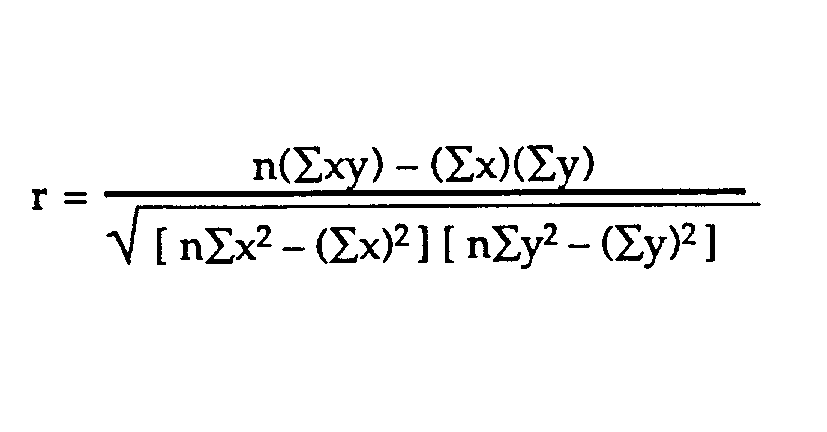
Range of R is -1 to 1

-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?

**Solution:**

Outliers are the data points which are distant than rest of the datapoints.

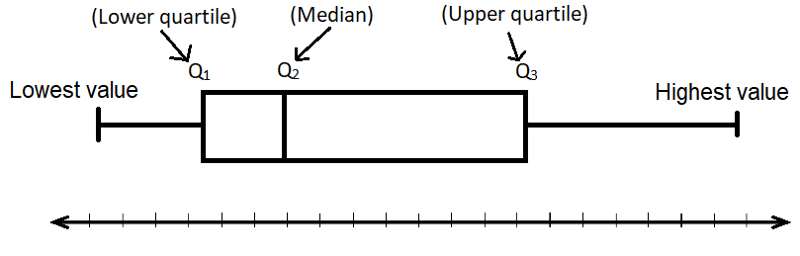
Following methods are used to detect and remove outliers:

1. **Box and Wisker plot**
2. **Grubb’s Test**
3. **Z-Score**
4. **IQR**

**1.Box-Wisker Plot**:

It is the graphical method to detect outliers.

There are basically five components of Box-Plot:



1. **Median Value**  
   Value that falls halfway between a set of values arranged in an ascending or descending order. When the set contains an odd number of values, the median value is exactly in middle. If the number of values is even, the median is computed by averaging the two numbers closest to the middle.
2. **Lower Quartile(Q1)or(Q25)**  
   The lower quartile is also known as the first quartile, splits the lower 25% of the data. Quartiles are three points that divide the data set into four equal groups. Each group represents the one-fourth of the data set. The lower quartile is the median of the lower half.
3. **Upper Quartile(Q3) or (Q75)**  
   Upper quartile is also known as the third quartile. It splits lowest 75% (or highest 25%) of data. It can be also seen as the middle value of the upper half.
4. **Interquartile Range(Q3-Q1)**  
   The Interquartile range is from Q1 to Q3. It is the difference between the lower quartile and upper quartile. The IQR is often seen as a better measure of a spread than the range (highest value-lowest value) as it is not affected by outliers.
5. **Upper Whisker**  
   It is k\*IQR offset above Q3, it shows the highest limit of boxplot. i.e Higest acceptable value. Anything beyond that can be termed as outlier.
6. **Lower Whisker**

It is k\*IQR offset below Q1, it shows the lowest limit of boxplot. i.e lowest acceptable value. Anything beyond that can be termed as outlier.

**2.Grubb’s Test:**

It is a statistical test to detect single outlier in a normally distributed dataset.

It works only if data distribution is normal.

It is feasible if you are sure that dataset does not have multiple outliers.

In Grubb’s test we use two statistics:

1. G-test statisctic
2. G-critical statisctic

G-test statisctic is calculated as follows:

The Grubbs’ test statistic for a [two-tailed test](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/) is:  
[grubb's test statistic](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/05/grubbs-test-statistic.png)  
  
  
Where:  
ȳ is the [sample mean](https://www.statisticshowto.datasciencecentral.com/sample-mean/),  
s = sample [standard deviation](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/standard-deviation/).

A **left-tailed test** uses the test statistic:  
[grubb's 2](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/05/grubbs-2.png)  
  
  
Where Ymin is the minimum value.

For a **right-tailed test**, use:  
[grubbs' test 3](https://www.statisticshowto.datasciencecentral.com/wp-content/uploads/2016/05/grubbs-3.png)

Where Ymin is the maximum value.

Then we have to find G-critical using critical value table and compare them

Compare your G test statistic to the G critical value:  
**Gtest < Gcritical**: keep the point in the data set; it is **not an outlier**.  
**Gtest > Gcritical:** **reject the point** as an outlier.

**3.Z-score:**

Z-score is a signed integer multiple of SD above or below the mean. It is a measure of variance in the dataset.

It is calculated as :

**Z = Xi – Xmean / SD**

A threshold Z-score value is selected like( -3 to 3) anything lying above this can be termed as outliers.

### **Why box plots?**

1. **A clear summary**  
   A box plot is a highly visually effective way of viewing a clear summary of one or more sets of data. It is particularly useful for quickly summarizing and comparing different sets of results from different experiments
2. **Displays outliers**  
   A box plot is one of very few statistical graph methods that show outliers. There might be one outlier or multiple outliers within a set of data, which occurs both below and above the minimum and maximum data values
3. **The Box plot as an Indicator of Centrality**  
   The median line shows the center of data which can be fruitfull in guessing relative median among different variable.
4. **The Box plot as an indicator of the spread**The thickness of the box can give an idea about symmetry of data distribution.
5. **The Box plot as an indicator of symmetry**The location of median line tell us about data distribution, e.g If line is towards lower half then data is right skewed and vice-versa.
6. **Independent of data distribution:**

The data distribution can be normal or not but it doesn’t effect the box plot which is not the case in Grubb’s Test.

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

**Solution:**

There are 3 method to impute missing values for the dataset

1. **Imputation with mean**

This strategy can be applied on a feature which has numeric data like the age of a person or the Salary. We can calculate the mean of the feature and replace it with the

missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns.

1. **Imputation with median**

Imputaion with median is same as that of mean , where it imputes the missing value with the median of dataset but it has one advantage over mean imputation and that is it more resilient for the outlliers.

1. **KNN Imputation :**

KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values.

To choose which is the best method for imputation we undergo a trial and error method :

1. **We make a certain datapoint null purposely and note it’s value.**
2. **Then perform The above methods one-by-one .**
3. **Note the value of the datapoint which was made null on purpose.**
4. **Compare the obtain values by each method.**
5. **Choose the method whose value is closest to true value .**