Correlation

Pearson Corr coeff

f_1 → y → $-1 \leftrightarrow +1$

↑ ↑

Input Output

-1 ↓ Strong inverse linear relationship

$+1$ → Strong positive linear relationship

f_1 f_2 f_3 ----- f_n y

$f_1 \rightarrow y$

$f_2 \rightarrow y$

$f_3 \rightarrow y$

⋮

$f_n \rightarrow y$

Catt off → 0.3

−0.3

↓

feature X

$f_1 - f_2$ $f_1 - f_3$ $f_1 - f_4$

multicollinear → $f_1 - (f_2) \rightarrow 0.9$ → drop

$f_1 - (f_3) \rightarrow 0.85$

Disadvantage

1. **Linearity Assumption** :- Correlation measures the linear relationship between two variables. It does not capture non-linear relationships well. If a relationship is non-linear, the correlation coefficient can be misleading.

2. **Doesn't Capture Complex Relationship** :- Correlationship only measures the relationship between two variables at a time. It may not capture complex relationships involving more than two variables. $f_1 \rightarrow f_2$ $\boxed{f_1\ f_2\ f_3} \rightarrow Y$

3. **Threshold Determination**: Just like variance ~~this~~ threshold, defining what level of correlation is considered "high" can be subjective and may vary depending on the specific problem or data set. $0.95 \rightarrow 0.9 \rightarrow 0.8$

4. **Sensitive to Outliers** :- Correlation is sensitive to outliers. A few extreme values can correlation coefficient.

4. Anova

1. Numerical → Y
   ↓f
   Categorical
   ↓
   more than 2 classes

2. Numerical → Y
   ↓f
   ↓
   numerical

Source of variation	Sum of squares (SS)	Degree of freedom (d.f)		
Between samples or categories	$n_1(\bar{x}_1 - \bar{\bar{x}})^2 + \dots n_k(\bar{x}_k - \bar{\bar{x}})^2$	$(k-1)$	$\frac{SS\ between}{(k-1)}$	$\frac{MS\ bet^n}{MS\ within}$
within samples or categories	$\sum (x_{i1} - \bar{x}_1)^2 + \dots \sum (x_{ik} - \bar{x}_k)^2$ $i = 1, 2, 3 \dots$	$(n-k)$	$\frac{SS\ within}{(n-k)}$	
Total	$\sum (x_{ij} - \bar{\bar{x}})^2$ $i, j = 1, 2, 3$	$(n-1)$		

n = rows

k = categories

f - statistic → p value → f ↔ y

f_1	Y
1	S
2	L
3	W
4	S
⋮	⋮
$\bar{\bar{X}}$	

↳ grand mean

⇒

S	L	W
1	2	3
4	5	7
3	-6	-8
$\bar{X}_S$	$\bar{X}_L$	$\bar{X}_W$

SS_W

$(1-\bar{X}_S)^2 + (4-\bar{X}_S)^2 + (3-\bar{X}_S)^2$

$+ (2-\bar{X}_L)^2 + (5-\bar{X}_S)^2 + (6-\bar{X}_S)^2$

$+ (3-\bar{X}_W)^2 + (7-\bar{X}_W)^2 + (8-\bar{X}_W)^2$

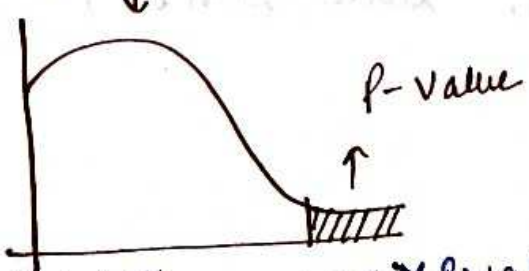
after calculate $SS_W = 50$

SS_b $\quad 1(\bar{X}_S - \bar{\bar{X}})^2 + 4(\bar{X}_S - \bar{\bar{X}})^2 + 3(\bar{X}_S - \bar{\bar{X}})^2$

$+ 2(\bar{X}_L - \bar{\bar{X}})^2 + 3(\bar{X}_L - \bar{\bar{X}}) + 6(\bar{X}_S - \bar{\bar{X}})^2$

$$\frac{\frac{SS_W}{k-1}}{\frac{SS_{btn}}{n-k}} = \text{f-ratio}$$

↳ dist → f-dist



$f_1 \to Y \Rightarrow$ No relation

reject Null Hypothesis → 0.05 > Pvalue

Disadvantage

1. Assumption of Normality : ANOVA assumes that the data for each group follow a normal distribution. This assumption may not hold true for all datasets, especially those with skewed distribution.

2. Assumption of Huniogeneity Of Variance :- ANOVA assumes that the variances of the different groups are equal. This is the assumption of homogeneity of variance (also known as homoscedasticity). If this assumption is violeted, it may lead to incorrect results.

3] Independence of Observations : ANOVA assumes that the observations are independent of each other. This might not be the case in datasets where observations are related (eg, time series, nested data).

4.) Effect of Outliers ?- ANOVA is sensitive to outliers. A single outlier can significantly affect the f-statistic leading to a potentially erroneous conclusion.

5] Doesn't Account for Interactions : Just like other univariate feature selection methods. ANOVA does not consider interactions bet^n features.

f_1 f_2

longitude latitude

Chi - Square

f_1 → Category

Y → Category

$f_1 \rightarrow$ Sex → form Contigency table (observed value)

↳ frequency table

relation

↳ chi-square

	0	1
M	468	109
F	81	233

Expected Value ideal

	0	1	
M	81	468	549
F	233	109	342 ⇒ Margin
Total	314	577	891

$(M,0) = \frac{314 \times 549}{891}$

$(F,0) = \frac{314 \times 342}{891}$

$(F,1) = \frac{342 \times 577}{891}$

$(M,1) = \frac{577 \times 549}{891}$

epected value

	0	1
M	193	355
F	120	221

diff ↓

crazy

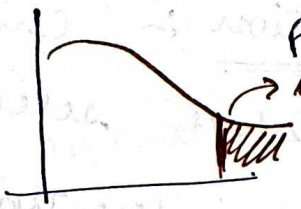
↳ large diff

↳ $f_i \rightarrow Y$ → change in f_1 and find variance Y

Chi-squared stat $= \left(\frac{\text{Observed value} - \text{expected value}}{\text{expected value}}\right)^2$

$= \frac{(468-493)^2}{193} + \frac{(109-355)^2}{355} + \left(\frac{81-120}{120}\right)^2 + \left(\frac{233-221}{100}\right)^2$

let Chi-squared stat = 15



← reject Null Hypo

Pvalue ↓↓ = Crazy ↑ & diff

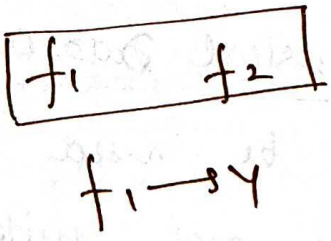
Disadvantage

1. Categorical Data Only :- The Chi-square test can only be used with categorical variables. It is not suitable for continous variables unless they have been discretized into categories, which can lead to loss of information.

2. Independence of Observations: The chi-square test assumes that the observation are independent of each other. This might not be the case in datasets where observations are related (eg. time series data, nested data)

3. Sufficient Sample size :- Chi-square test requires a sufficiently large sample size. The result may not be reliable. If the sample size is too small or if the frequency count in any category is too low (typically less than 5)

4. No variable interaction :- Chi-square test, like other univariate feature selection methods, does not consider interaction between features. It might miss out identifying important feature that are significant in combination with other features.



Advantage and disadvantage

Advantage

1. Simplicity: filter methods are generally straight-forward and easy feature and selectly the top features based on this statistics.

2. Speed:- These methods are usually computationally efficient. Because the evaluate because they evaluate each feature independently, they can be much faster than wrapper methods or embedded methods, which need to train a model to evaluate freature importance.

3. Scalablity: Filter methods can handle a large number of features efficitively because they dont involve any learning methods. This makes them suitable for higher dimensional datasets.

4. Pre-processing step: They can serve as a pre-processing step for other feature selection methods. For instance, you could use a filter

method to remove irrelevant features before applying a more computationally expensive method, such as a wrapper method.

Disadvantage

1. **Lack of feature Interaction** :- Filter method treat each feature individual and hence do not consider the interaction between feature. They might miss out on identifying important features that doesn't appear significant individual but are significant in combination with other feature.

2. **Model Agnostic** :- Filter method are agnostic to the machine learning model that will be used for the prediction. This means that the selected feature might not necessarily contribute to the accuracy of the specific model you want to use.

3. **Statistical Measures Limitation** :- The statistical measure used in these methods have their own limitation. For example,

correlation is the measure of the linear relationship and might not capture non-linear relationship effectively. Similarly, variance based methods might keep feature with high variance but low predictive power.

4. Threshold determination :- For some methods, determining the threshold to select features can be a bit subjective. For example, what constitutes "low" variance or "high" correlation might differ depending on the context or the specific dataset.