Train In Data

Chi-square

Chi-square

Chi-square is a statistical test, best suited to determine a difference between expected frequencies and observed frequencies in 1 or more categories of a contingency table.

Chi-square – contingency table

- **Null hypothesis**: gender does not indicate survival.
- **Count** number of people from each gender who survived and didn't.
- **Divide** each count by the total people from each gender
- **Expected frequency**: value expected should there be no difference between gender.

	Male	Female	Expected frequency
Survived = 1	161	337	498
Survived = 0	681	127	808
Total column	842	464	1306

	Male	Female	Expected frequency
Survived = 1	0.19	0.73	0.38
Survived = 0	0.81	0.27	0.62
Total column	0.19	0.73	0.38

Chi-square – contingency table

- **Expected frequency**: value expected should there be no difference between gender.

- Male and Female frequencies are different from the expected ones
  - The feature is useful to predict survival

	Male	Female	Expected frequency
Survived = 1	161	337	498
Survived = 0	681	127	808
Total column	842	464	1306

	Male	Female	Expected frequency
Survived = 1	0.19	0.73	0.38
Survived = 0	0.81	0.27	0.62
Total column	0.19	0.73	0.38

Chi-square – contingency table

- **Calculate the statistic:**

$$X2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- **Calculate the degrees of freedom:**

(rows -1)(columns-1) = 1

- **Compare the statistic against a known distribution ➔ chi-square**

	Male	Female	Expected frequency
Survived = 1	161	337	498
Survived = 0	681	127	808
Total column	842	464	1306

	Male	Female	Expected frequency
Survived = 1	0.19	0.73	0.38
Survived = 0	0.81	0.27	0.62
Total column	0.19	0.73	0.38

Chi-square

- Suited for categorical variables.
- Target should be categorical.
- Variable values should be non-negative, and typically Boolean, frequencies, or counts.
- It compares observed distribution of class among the different labels against the expected one, would there be no labels.

Chi-square: Scikit-learn

- **Chi2**: ranks features ➔ smallest the p-value biggest importance
  - Spin implementation of scipy.stats.chisquare
  - https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html
- **SelectKBest**: select best k features
- **SelectPercentile**: select features in top percentile

Train In Data

THANK YOU

www.trainindata.com