Statistical Tests and Evaluation Metrics

Metric/Test	Purpose	Key Value	Significance Interpretation
t-test	Compare means between two groups	p-value	Low p-value (< 0.05) indicates significant difference between groups
ANOVA (Analysis of Variance)	Compare means among three or more groups	p-value, F-statistic	Low p-value (< 0.05) indicates significant difference among groups
Chi-square test	Test for association between categorical variables	p-value, Chi-square	Low p-value (< 0.05) indicates significant association
Correlation (Pearson)	Measure linear relationship between two continuous variables	r-value, p-value	High absolute r-value (close to 1) and low p-value (< 0.05) indicate strong correlation
Regression Analysis	Predict value of dependent variable based on independent variable(s)	p-value, R-squared	Low p-value (< 0.05) for coefficients indicates significant predictors; high R-squared indicates good model fit
Logistic Regression	Predict probability of binary outcome	p-value, Odds Ratio	Low p-value (< 0.05) for coefficients indicates significant predictors; odds ratio > 1 or < 1 indicates impact direction

Mann-Whitney U test	Compare medians between two independent groups	p-value, U statistic	Low p-value (< 0.05) indicates significant difference between groups
Wilcoxon Signed-Rank test	Compare medians between two related groups	p-value, W statistic	Low p-value (< 0.05) indicates significant difference between paired samples
Kruskal-Wallis test	Compare medians among three or more groups	p-value, H statistic	Low p-value (< 0.05) indicates significant difference among groups
Fisher's Exact Test	Test for association between small sample categorical variables	p-value	Low p-value (< 0.05) indicates significant association
Cox Proportional Hazards Model	Assess effect of variables on survival time	p-value, Hazard Ratio	Low p-value (< 0.05) for coefficients indicates significant predictors; hazard ratio > 1 or < 1 indicates impact direction
Jaccard Index	Measure similarity between two sets	Jaccard Index	High Jaccard Index (close to 1) indicates high similarity
F1-score	Measure test's accuracy (harmonic mean of precision and recall)	F1-score	High F1-score (close to 1) indicates better model performance
Log Loss	Measure performance of a classification model (probability estimates)	Log Loss	Low Log Loss indicates better model performance

Accuracy	Proportion of correctly classified instances	Accuracy	High accuracy indicates better model performance
Precision	Proportion of true positive results among the predicted positives	Precision	High precision indicates better model performance
Recall (Sensitivity)	Proportion of true positive results among the actual positives	Recall	High recall indicates better model performance
ROC-AUC (Receiver Operating Characteristic - Area Under Curve)	Measure performance of binary classification models	AUC	High AUC (close to 1) indicates better model performance
Confusion Matrix	Summarize performance of classification algorithm	TP, FP, TN, FN	High TP and TN, low FP and FN indicate better model performance
K-fold Cross-Validation	Assess model performance by dividing data into k subsets and rotating validation	Mean score, standard deviation	Lower standard deviation indicates more stable model performance
Leave-One-Out Cross-Validation (LOOCV)	Assess model performance by using one observation as the validation set in each iteration	Mean score, standard deviation	Lower standard deviation indicates more stable model performance
Bootstrapping	Estimate the distribution of a statistic by sampling with replacement	Mean score, confidence intervals	Narrow confidence intervals indicate more precise estimates
Adjusted R-squared	Adjust R-squared for the number of predictors in the model	Adjusted R-squared	Higher Adjusted R-squared indicates better model fit

BIC/AIC (Bayesian/ Akaike Information Criterion)	Evaluate model fit with a penalty for complexity	BIC/AIC	Lower BIC/AIC indicates better model fit
Silhouette Score	Evaluate clustering performance	Silhouette Score	High silhouette score (close to 1) indicates well-defined clusters
Davies-Bouldin Index	Evaluate clustering performance	Davies-Bouldin Index	Lower Davies-Bouldin Index indicates better clustering
Inertia (within-cluster sum of squares)	Evaluate clustering performance	Inertia	Lower inertia indicates better clustering
Rand Index	Measure similarity between two clustering results	Rand Index	High Rand Index (close to 1) indicates better clustering
Mutual Information	Measure dependency between two variables	Mutual Information	High mutual information indicates stronger dependency
Purity Score	Evaluate clustering performance	Purity Score	High purity score indicates better clustering

Key Terms

1. **p-value:** Probability that the observed data would occur by chance. A low p-value (< 0.05) typically indicates statistical significance.
2. **r-value:** Correlation coefficient representing the strength and direction of a linear relationship.
3. **F-statistic:** Ratio used in ANOVA to determine the significance of group differences.
4. **Chi-square:** Statistic used to measure the association between categorical variables.
5. **R-squared:** Proportion of variance explained by the model in regression analysis.
6. **Odds Ratio:** Measure of association between an exposure and an outcome in logistic regression.
7. **U statistic:** Value calculated in the Mann-Whitney U test to compare medians.
8. **W statistic:** Value calculated in the Wilcoxon Signed-Rank test for paired samples.
9. **H statistic:** Value calculated in the Kruskal-Wallis test to compare medians among groups.
10. **Hazard Ratio:** Measure of effect in survival analysis; values > 1 or < 1 indicate increased or decreased hazard, respectively.
11. **Jaccard Index:** Measures similarity between two sets; defined as the size of the intersection divided by the size of the union of the sets.
12. **F1-score:** Harmonic mean of precision and recall, providing a balance between the two metrics; useful for imbalanced datasets.
13. **Log Loss (Logarithmic Loss):** Measures the performance of a classification model where the prediction is a probability value between 0 and 1; lower values indicate better performance.
14. **True Positives (TP):** Correctly predicted positive instances.
15. **False Positives (FP):** Incorrectly predicted positive instances.
16. **True Negatives (TN):** Correctly predicted negative instances.
17. **False Negatives (FN):** Incorrectly predicted negative instances.

18. **K-fold Cross-Validation:** A technique to assess model performance by splitting data into k subsets and using one subset for validation while the remaining k-1 subsets are used for training, repeated k times.
19. **Leave-One-Out Cross-Validation (LOOCV):** A technique where each observation is used once as a validation set, and the rest are used as the training set.
20. **Bootstrapping:** A resampling method to estimate the distribution of a statistic by sampling with replacement.
21. **Silhouette Score:** Measures how similar an object is to its own cluster compared to other clusters; ranges from -1 to 1.
22. **Davies-Bouldin Index:** Measures the average similarity ratio of each cluster with the one most similar to it; lower values indicate better clustering.
23. **Inertia:** Sum of squared distances of samples to their closest cluster center.
24. **Rand Index:** Measures the similarity between two data clusterings.
25. **Mutual Information:** Measures the amount of information obtained about one variable through the other variable.
26. **Purity Score:** Measures the extent to which clusters contain a single class.