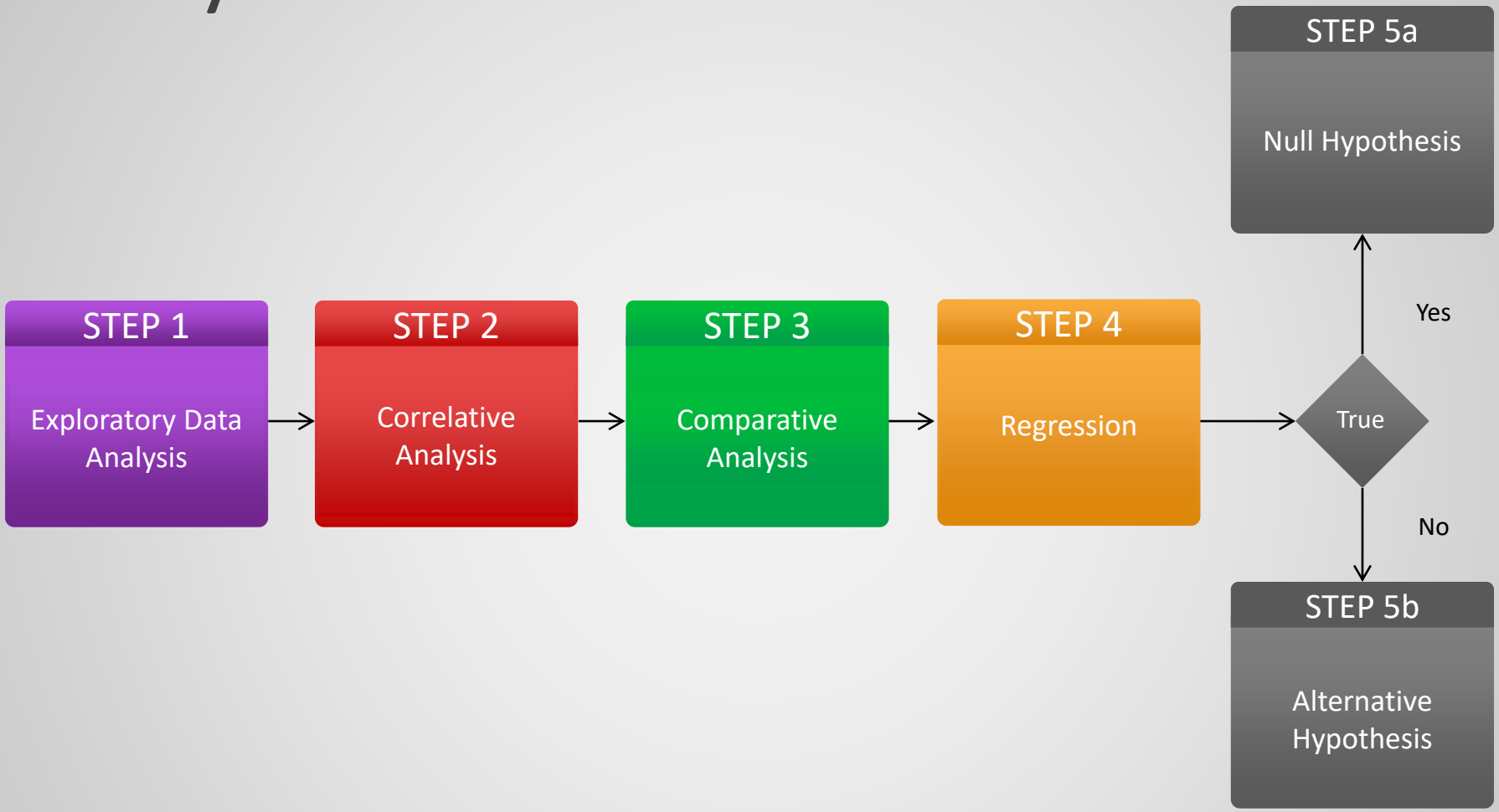
Data Analysis Process

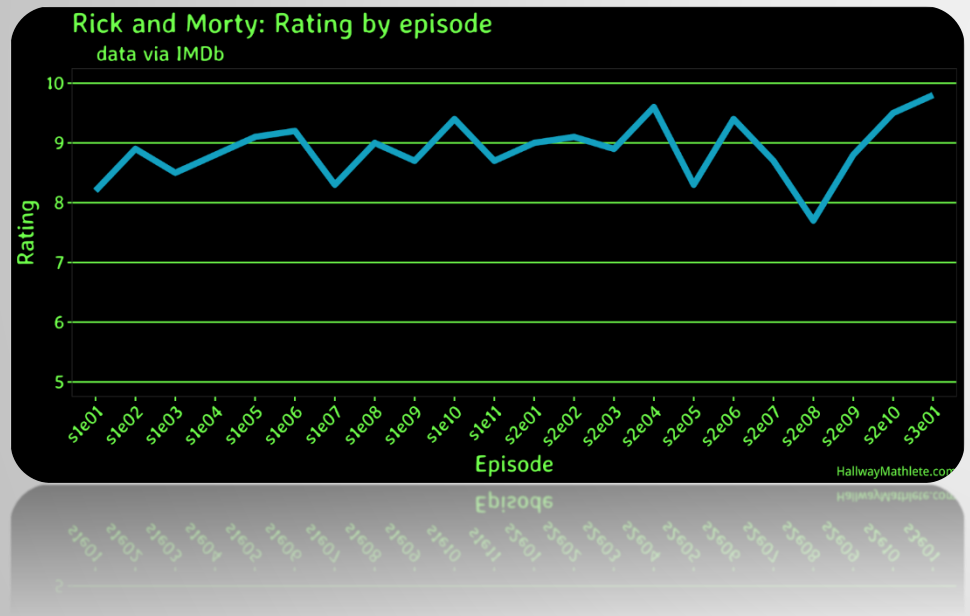


Danielle Lynch
Level Data Analytics
Northeastern University
February 2019

Descriptive Analysis (also EDA)

STEP 1

Exploratory Data Analysis (EDA)



Descriptive Stats

Categorical

- Show frequencies
- Minimum
- Maximum
- Percentages

Heat map
Histogram
Bar chart
Scatterplot
Line graph

Numerical

- Summary statistics
- Assess normality
- Identify outliers
- Central Tendency

Heat map
Distribution plot
Box plot

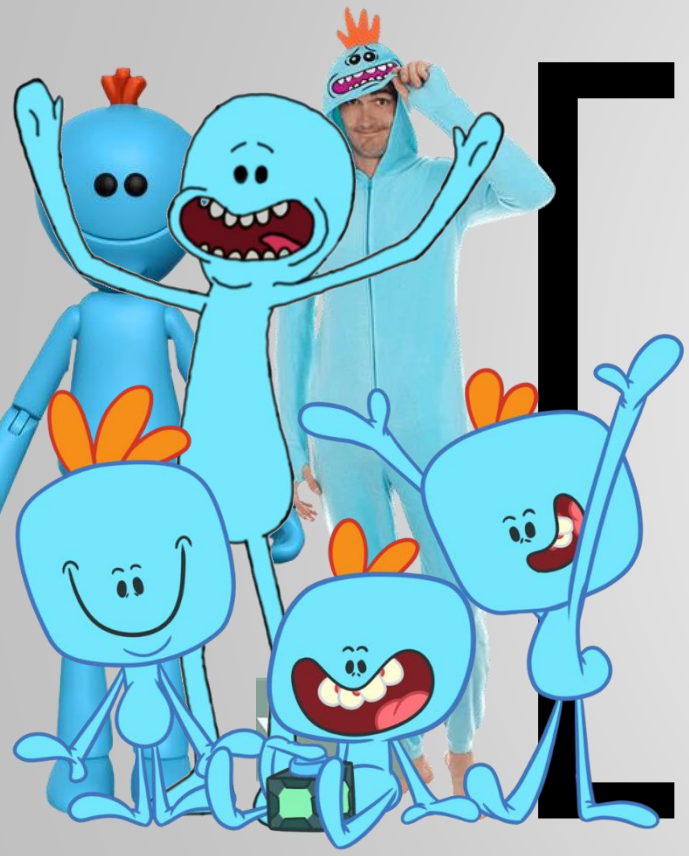
Missing Data

- By variable
- By element
- 5% threshold
- Missing data handling

Missing Data Handling

- Missing data
  - Deletion
    - Listwise deletion (complete cases)
    - Pairwise deletion
    - Dropping variables
  - Imputation
    - Time-series
      - With trend & seasonality — Seasonal adjustment + interpolation
      - With trend but without seasonality — Linear interpolation
      - Without trend & seasonality — Mean, median, mode, random sample imputation
    - General
      - Categorical — Make NA as level, multiple imputation, logistic regression
      - Continuous — Mean, median, mode, multiple imputation, linear regression

Correlative and Comparative Analysis



STEP 2

Correlative Analysis

STEP 3

Comparative Analysis

Difference/Similarity

- Ranking
- Clustering analysis
- Magnitude
- Direction of difference

Parallel coordinates plot
Difference plot
Multidimensional scaling
Dendogram
Cluster plot

Correlation

- Association testing
- Principle component analysis
- Factor analysis
- Correlation coefficient

Correlation plot
Scree plot
Factor score plot
Cluster bar chart
Q-Q plot

Comparison

- “Goodness of fit”
- T-Distributions
- Confidence intervals
- Comparison of means
- Analysis of variance
- Significance testing

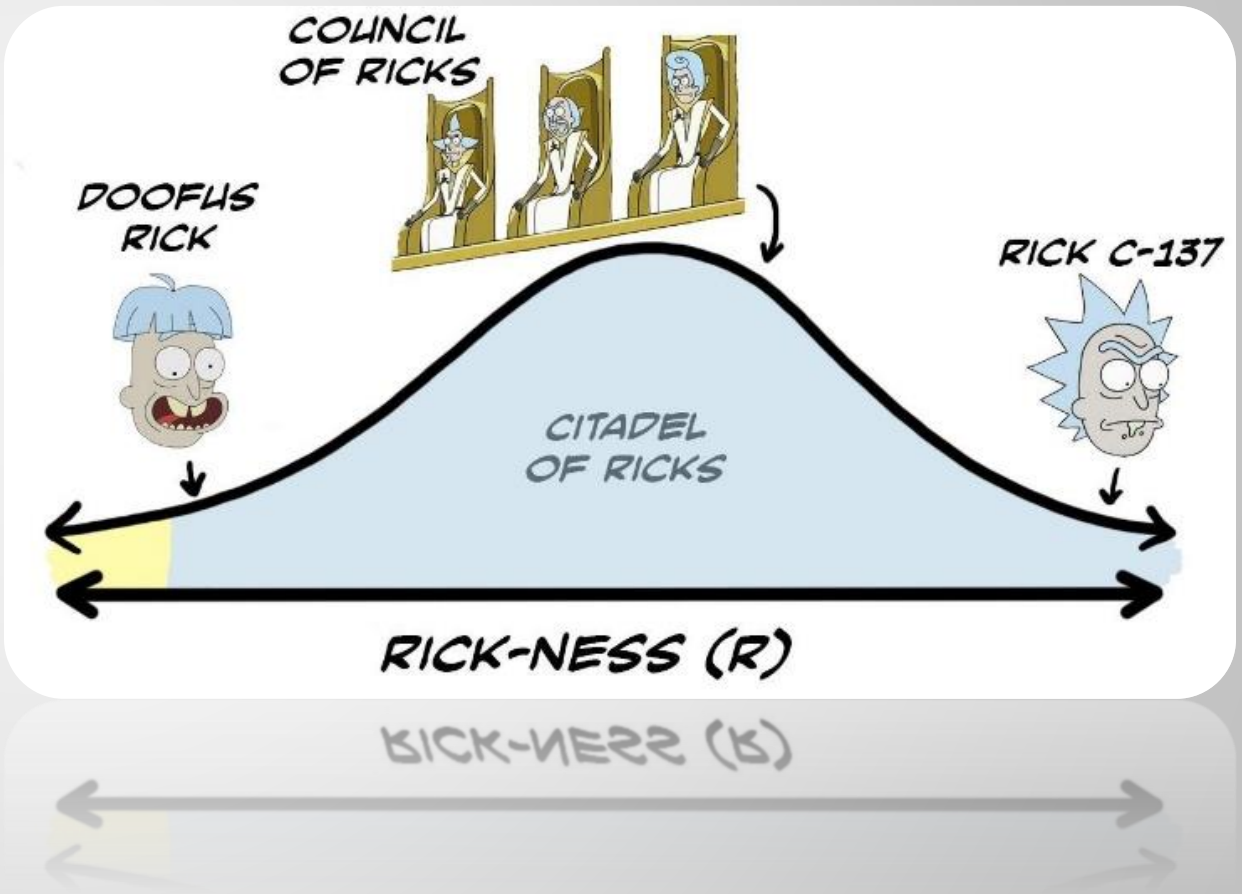
Distribution plot
ANOVA plot
Fitted variance plot

Bivariate Inferential Techniques

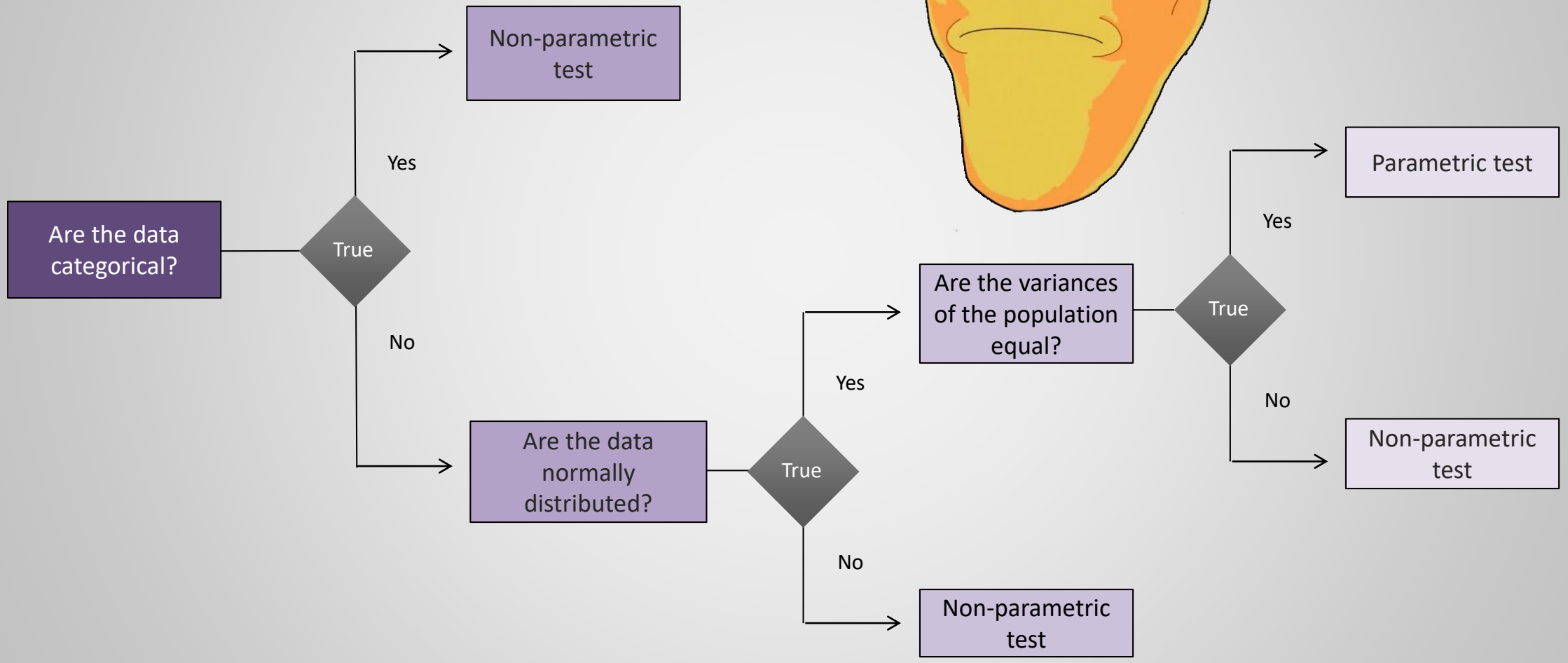
Categorical by Categorical	Continuous by Categorical	Continuous by Continuous
CROSSTABS	MEANS or EXPLORE	CORRELATIONS
Chi-Square Statistic	T-test or One-way ANOVA	Correlations or Regression
Clustered Bar Chart	Box Plot or Error Bar Chart	Scatterplot

Parametric v. Non-Parametric Tests

	Parametric	Non-Parametric
Assumed distribution	Normal	Any
Assumed variance	Homogenous	Any
Typical data	Ratio or Interval	Ordinal or Nominal
Usual central measure	Mean	Median
Data set relationships	Samples are independent	Any
Benefits	Can draw more conclusions Usually the better, more powerful choice	Simplicity Less affected by outliers Can tell if a difference exists but not how big it is
Large samples (don't worry!)	Extremely robust even if the distribution is non-normal	Still very powerful even if the distribution is normal
Small samples (be careful!)	P value may be inaccurate if the distribution is non-normal	P value will be too high if the distribution is normal

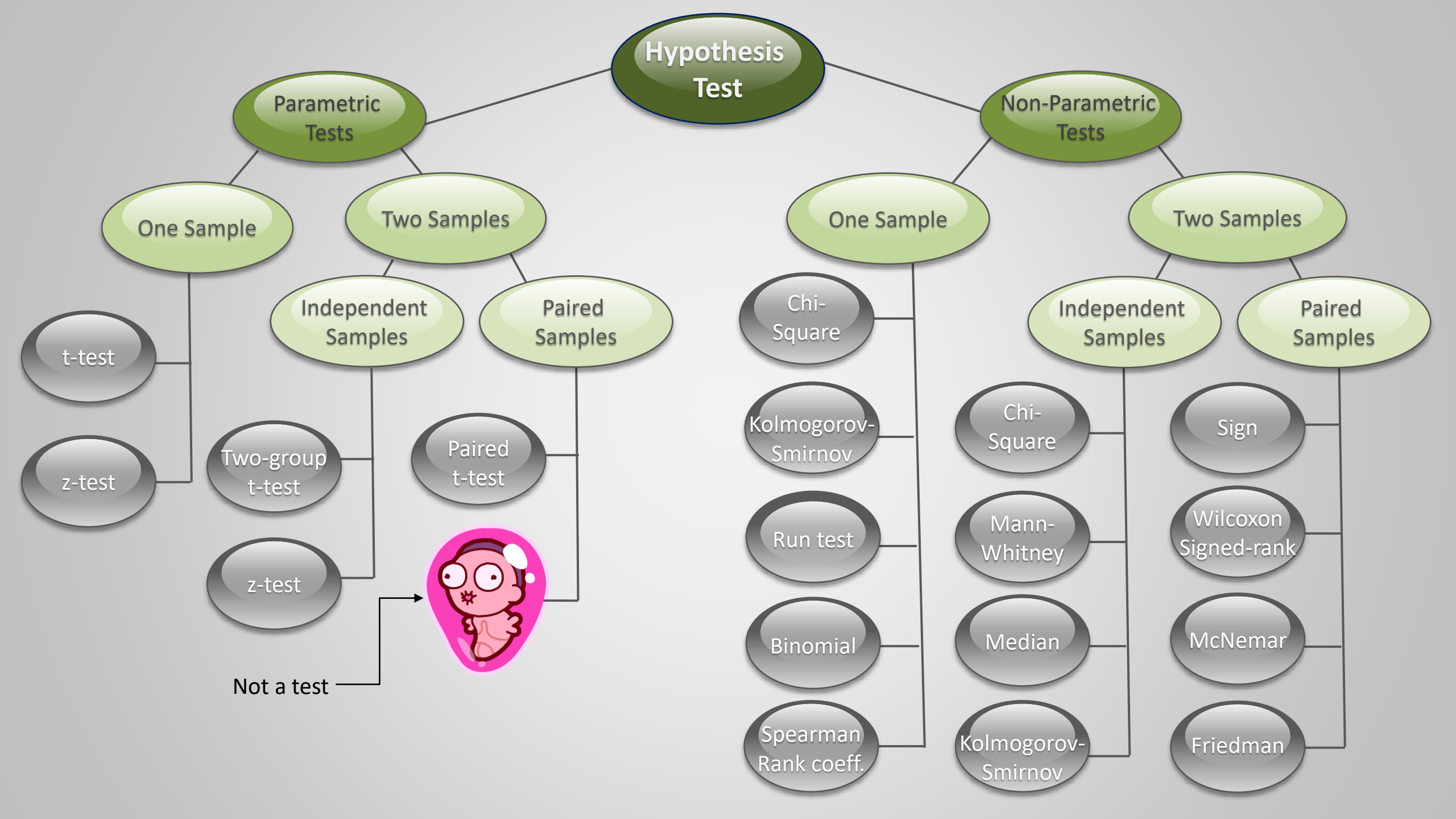


Parametric v. Non-Parametric Decision Tree



Parametric v. Non-Parametric Tests

Analysis Type	Parametric Procedure	Non-Parametric Procedure	BONUS INFO: Binomial (Two Possible Outcomes)
Describe one group.	Mean, SD	Median, interquartile rage	Proportion
One group, variance unknown. Determines whether the sample mean is statistically different from a known or hypothesized population mean.	One-sample t-test	One-sample Sign One-sample Wilcoxon	Chi-square or Binomial test
One group, variance known. Determines whether the sample mean is statistically different from a known or hypothesized population mean.	One-sample z-test	One-sample Sign One-sample Wilcoxon	
Two unpaired (independent) groups. Determines whether two population means are equal or unequal.	Two-sample t-test	Mann-Whitney U test; Wilcoxon rank-sum test	Fisher's test (chi-square for large samples)
Two paired (dependent) groups. Determines whether two population means are equal or unequal.	Paired sample t-test	Wilcoxon test	McNemar's test
Compares three or more means of unmatched groups.	One-way ANOVA	Kruskal-Wallis test Mood's Median test	Chi-square test
Compares three or more means of matched groups.	Two-way ANOVA or Repeated-measures ANOVA	Friedman's test	Cochrane Q
Quantifies the association between two continuous variables.	Pearson correlation	Spearman correlation	Contingency coefficients
Two unpaired (independent) groups. Determines whether the two variances are equal or unequal.	F-test for two variances		
Determines whether the observed and expected variances are equal or unequal.		Chi-squared test for single variance	


Hypothesis Test
Parametric Tests
Non-Parametric Tests
One Sample
Two Samples
One Sample
Two Samples
t-test
z-test
Independent Samples
Paired Samples
Two-group t-test
z-test
Paired t-test
Not a test
Chi-Square
Kolmogorov-Smirnov
Run test
Binomial
Spearman Rank coeff.
Independent Samples
Paired Samples
Chi-Square
Mann-Whitney
Median
Kolmogorov-Smirnov
Sign
Wilcoxon Signed-rank
McNemar
Friedman

Predictive Analysis (Regression)

STEP 4

Regression

Regression Analysis Process:

1. Define dependent and independent variables
2. Hypothesize nature of relationship between variables
3. Estimate unknown model coefficients
4. Test the model and adjust the form of the estimated regression equation as needed

Regression Types

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Quantile Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- Principal Component Regression
- Partial Least Square Regression
- Support Vector Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Zero-Inflated Poisson Regression
- Quasi-Poisson Regression
- Cox Regression

Considerations

- Outliers
- Residuals
- Multicollinearity
- Heterodasticity
- Under/Overfitting

How to know if a model is the best fit?

Statistic	Criterion
R^2 and Adjusted R^2	Higher the better (> 0.70)
F-statistic	Higher the better
Standard Error	Closer to zero the better
t-statistic	Higher the better
Akaike information criterion (AIC)	Lower the better (when comparing models)
Bayesian information criterion (BIC)	Lower the better (when comparing models)
Mallows cp	Should be close to the number of predictors in model
MAPE (Mean absolute percentage error)	Lower the better
MSE (Mean squared error)	Lower the better
Min-Max Accuracy => mean(min(actual, predicted)/max(actual, predicted))	Higher the better
p-value	Very small (p < alpha) to reject the null hypothesis.
Chi-square test statistic	Lower the better
z-score	Either very high or very low to reject the null hypothesis.
Residual standard error (RSE)	Closer to zero the better
Stars (*) in the summary table	The more the stars beside a variable's p-value, the more significant the variable.
Deviance	Lower the better (a value of zero only happens if the model is a perfect fit)

Linear Regression

The simplest form of regression. It is a technique in which the dependent variable is continuous in nature.

Assumptions

- The relationship between the dependent and independent variables is linear.
- Observations must be independent.
- No significant outliers.
- No heteroscedasticity.
- No multicollinearity or auto-correlation.

Limitations

- Often has low bias but high variance.
- Only looks at the mean of the dependent variable.
- Ascertains relationships but not the underlying causal mechanism.

Key Questions

- ✓ Are there at least 10 to 20 times as many observations as variables?
- ✓ Is the coefficient of determination (R^2) high?
- ✓ In the case of multiple predictors, is the variance inflation factor low?
- ✓ Are errors (residuals) normally distributed?
- ✓ Do errors (residuals) have a constant variance?
- ✓ Are errors (residuals) independent of each other?

Linear model

lm(y ~ x)

Polynomial Regression

It is a technique to fit a nonlinear equation by taking polynomial functions of the independent variable.

Assumptions

- The relationship between the dependent and independent variables is linear or curvilinear.
- Observations must be independent.
- No significant outliers.
- No heteroscedasticity.
- No multicollinearity or auto-correlation.

Limitations

- High order polynomials (n > 4) may lead to over-fitting.
- Strong sensitivity to outliers.
- Fewer model validation tools.
- Useless outside the range of observed data, do not try to extend predictions beyond this range.

Key Questions

- ✓ Does a simple scatter plot reveal a curvilinear relationship?
- ✓ Does a scatterplot of residuals show patches of residuals in the middle?
- ✓ Is the coefficient of determination (R^2) high?
- ✓ Are errors (residuals) normally distributed?
- ✓ Do errors (residuals) have a constant variance?
- ✓ Are errors (residuals) independent of each other?

Quadratic model

lm(y ~ poly(x, 2))

Cubic model

lm(y ~ poly(x, 3))

Quantile Regression

An extension of the linear regression model that predicts the entire conditional distribution of the response, not just the mean.

Assumptions

- The target variable needs to be a continuous variable.
- Does not assume normality or linearity.

Limitations

- Is computationally intensive.
- Needs sufficient data.
- Can not take account of time–varying covariates.

Key Questions

- ✓ Are there significant outliers?
- ✓ Are the data highly skewed?
- ✓ Does heterodasticity exist in the data?

Quantile model

rq(y ~ x, tau = 0.xx)

Logistic Regression

It is a technique to model the probabilities of success of a binomial outcome.

Assumptions

- Independent variables can be continuous or binary.
- The dependent variable must be binary.
- Sample observations are independent.
- Little to no multicollinearity or auto-correlation amongst independent variables.

Limitations

- Only useful if all the relevant independent variables have already been identified.
- Only concerned about the probability of outcome for the dependent variable (success or failure).
- Typically requires a large sample size.

Key Questions

- ✓ Are the predictor values normally distributed?
- ✓ Is the area under the ROC curve large?
- ✓ Is the AUC-ROC score high?
- ✓ Is there a large difference between null and residual deviance?
- ✓ Is deviance low using anova()?
- ✓ Is the probability score high?

Logistic model

```
glm(y ~ x,
fam = “binomial”)
```

Poisson Regression

It is the simplest regression model used when the dependent variable has count data of rare events.

.

Assumptions

- The variance of the observations is equal to the mean (equidispersion).
- Counts must be positive integers.
- Observations must be independent.
- On average, events occur randomly yet uniformly in time.

Limitations

- Sensitive to overdispersion and excess zeros.
- Failure to recognize equidispersion.
- Failure to recognize if the data are truncated.
- Failure to model the variance of date.

Key Questions

- ✓ Is the dispersion parameter approximately equal to 1?
- ✓ Is the residual deviance approximately equal to the degrees of freedom?
- ✓ Is deviance low using anova()?
- ✓ Do errors (residuals) follow a Poisson distribution?

Poisson model

```
glm(y ~ x,
fam = “poisson”)
```

Negative Binomial Regression

A type of Poisson regression that addresses overdispersion.

Assumptions

- The variance of the observations is greater than the mean (overdispersion).
- Counts must be positive integers.
- Observations must be independent.
- On average, events occur randomly yet uniformly in time.

Limitations

- Sensitive to excess zeros.
- Failure to recognize if the data are truncated.
- Failure to model the variance of date.

Key Questions

- ✓ Is the dispersion parameter > 1?
- ✓ Is the residual deviance approximately equal to the degrees of freedom?
- ✓ Is deviance low using anova()?
- ✓ Do errors (residuals) follow a Poisson distribution?

NB model

glm.nb(y ~ x)

Zero-Inflated Poisson Regression

A type of Poisson regression that addresses an overabundance of zero counts.

Assumptions

- The variance of the observations is equal to the mean (equidispersion).
- Counts must be positive integers.
- Observations must be independent.
- On average, events occur randomly yet uniformly in time.

Limitations

- Sensitive to excess overdispersion.
- Failure to recognize if the data are truncated.
- Failure to model the variance of date.

Key Questions

- ✓ Is the dispersion parameter approximately equal to 1?
- ✓ Is the residual deviance approximately equal to the degrees of freedom?
- ✓ Is deviance low using anova()?
- ✓ Do errors (residuals) follow a Poisson distribution?

ZIP model

pscl::zeroinfl(y ~ x)

Sometimes science is more art than science, Morty.