# Time Series

Extracts characteristics from time-sequenced data, which may exhibit the following characteristics:

- Stationarity - statistical properties such as mean, variance, and auto correlation are constant over time
- Trend - long-term rise or fall in values
- Seasonality - variations associated with specific calendar times, occurring at regular intervals less than a year
- Cyclicality - variations without a fixed time length, occurring in periods of greater or less than one year
- Autocorrelation - degree of linear similarity between current and lagged values

CV must account for the time aspect, such as for each fold $F_x$:

- Sliding Window - train $F_1$, test $F_2$, then train $F_2$, test $F_3$
- Forward Chain - train $F_1$, test $F_2$, then train $F_1, F_2$, test $F_3$

**Exponential Smoothing** - uses an exponentially decreasing weight to observations over time, and takes a moving average. The time $t$ output is $s_t = \alpha x_t + (1-\alpha)s_{t-1}$, where $0 < \alpha < 1$.

**Double Exponential Smoothing** - applies a recursive exponential filter to capture trends within a time series

$$s_t = \alpha x_t + (1-\alpha)(s_{t-1} + b_{t-1})$$
$$b_t = \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}$$

Triple exponential smoothing adds a third variable $\gamma$ that accounts for seasonality.

**ARIMA** - models time series using three parameters $(p, d, q)$:

- Autoregressive - the past $p$ values affect the next value
- Integrated - values are replaced with the difference between current and previous values, using the difference degree $d$ (0 for stationary data, and 1 for non-stationary)
- Moving Average - the number of lagged forecast errors and the size of the moving average window $q$

**SARIMA** - models seasonality through four additional seasonality-specific parameters: $P$, $D$, $Q$, and the season length $s$

**Prophet** - additive model that uses non-linear trends to account for multiple seasonalities such as yearly, weekly, and daily. Robust to missing data and handles outliers well. Can be represented as: $y(t) = g(t) + s(t) + h(t) + \epsilon(t)$, with four distinct components for the growth over time, seasonality, holiday effects, and error. This specification is similar to a generalized additive model.

**Generalized Additive Model** - combine predictive methods while preserving additivity across variables, in a form such as $y = \beta_0 + f_1(x_1) + \cdots + f_m(x_m)$, where functions can be non-linear. GAMs also provide regularized and interpretable solutions for regression and classification problems.

# Naive Bayes

Classifies data using the label with the highest conditional probability, given data $a$ and classes $c$. Naive because it assumes variables are independent.

**Bayes' Theorem** $P(c_i|a) = \frac{P(a|c_i)P(c_i)}{P(a)}$

**Gaussian Naive Bayes** - calculates conditional probability for continuous data by assuming a normal distribution

# Statistics

***p*-value** - probability that an effect could have occurred by chance. If less than the significance level $\alpha$, or if the test statistic is greater than the critical value, then reject the null.

**Type I Error** (False Positive $\alpha$) - rejecting a true null

**Type II Error** (False Negative $\beta$) - not rejecting a false null

Decreasing Type I Error causes an increase in Type II Error

**Confidence Level** (1 - $\alpha$) - probability of finding an effect that did not occur by chance and avoiding a Type I error

**Power** (1 - $\beta$) - probability of picking up on an effect that is present and avoiding a Type II Error

**Confidence Interval** - estimated interval that models the long-term frequency of capturing the true parameter value

***z*-test** - tests whether normally distributed population means are different, used when $n$ is large and variances are known

- z-score - the number of standard deviations between a data point $x$ and the mean $\rightarrow \frac{x-\mu}{\sigma}$

***t*-test** - used when population variances are unknown, and converges to the $z$-test when $n$ is large

- t-score - uses the standard error as an estimate for population variance $\rightarrow \frac{x-\mu}{s/\sqrt{n}}$

**Degrees of Freedom** - the number of independent (free) dimensions needed before the parameter estimate can be determined

**Chi-Square Tests** - measure differences between categorical variables, using $\chi^2 = \sum \frac{observed - expected}{expected}$ to test:

- Goodness of fit - if samples of one categorical variable match the population category expectations
- Independence - if being in one category is independent of another, based off two categories
- Homogeneity - if different subgroups come from the same population, based off a single category

**ANOVA** - analysis of variance, used to compare 3+ samples

- F-score - compares the ratio of explained and unexplained variance $\rightarrow \frac{\text{between group variance}}{\text{within group variance}}$

**Conditional Probability** $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

If $A$ and $B$ are independent, then $P(A \cap B) = P(A)P(B)$. Note, events that are independent of themselves must have probability either 1 or 0.

**Union** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

**Mutually Exclusive** - events cannot happen simultaneously

**Expected Value** $E[X] = \sum x_i p_i$, with properties

- $E[X+Y] = E[X] + E[Y]$
- $E[XY] = E[X]E[Y]$ if $X$ and $Y$ are independent

**Variance** $\text{Var}(X) = E[X^2] - E[X]^2$, with properties

- $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\text{Cov}(X, Y)$
- $\text{Var}(aX \pm b) = a^2\text{Var}(X)$

**Covariance** - measures the direction of the joint linear relationship of two variables $\rightarrow \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

**Correlation** - normalizes covariance to provide both strength and direction of linear relationships $\rightarrow r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

Independent variables are uncorrelated, though the inverse is not necessarily true

# A/B Testing

Examines user experience through randomized tests with two variants. The typical steps are:

1. Determine the evaluation metric and experiment goals
2. Select a significance level $\alpha$ and power threshold 1 - $\beta$
3. Calculate the required sample size per variation
4. Randomly assign users into control and treatment groups
5. Measure and analyze results using the appropriate test

The required sample size depends on $\alpha$, $\beta$, and the MDE

**Minimum Detectable Effect** - the target relative minimum increase over the baseline that should be observed from a test

**Overall Evaluation Criterion** - quantitative measure of the test's objective, commonly used when short and long-term metrics have inverse relationships

**Multivariate Testing** - compares 3+ variants or combinations, but requires larger sample sizes

**Bonferroni Correction** - when conducting $n$ tests, run each test at the $\frac{\alpha}{n}$ significance level, which lowers the false positive rate of finding effects by chance

**Network Effects** - changes that occur due to effect spillover from other groups. To detect group interference:

1. Split the population into distinct clusters
2. Randomly assign half the clusters to the control and treatment groups $A_1$ and $B_1$
3. Randomize the other half at the user-level and assign to control and treatment groups $A_2$ and $B_2$
4. Intuitively, if there are network effects, then the tests will have different results

To account for network effects, randomize users based on time, cluster, or location

**Sequential Testing** - allows for early experiment stopping by drawing statistical borders based on the Type I Error rate. If the effect reaches a border, the test can be stopped. Used to combat *peeking* (preliminarily checking results of a test), which can inflate $p$-values and lead to incorrect conclusions.

**Cohort Analysis** - examines specific groups of users based on behavior or time and can help identify whether novelty or primacy effects are present

# Miscellaneous

**Shapley Values** - measures the marginal contribution of each variable in the output of a model, where the sum of all Shapley values equals the total value (prediction $-$ mean prediction)

**SHAP** - interpretable Shapley method that utilizes both global and local importance to model variable explainability

**Permutation** - order matters $\rightarrow \frac{n!}{(n-k)!} = {}^nP_k$

**Combination** - order doesn't matter $\rightarrow \frac{n!}{k!(n-k)!} = {}^nC_k = \binom{n}{k}$

**Left Skew** - Mean < Median $\leq$ Mode

**Right Skew** - Mean > Median $\geq$ Mode

**Probability vs Likelihood** - given a situation $\theta$ and observed outcomes $O$, probability is calculated as $P(O|\theta)$. However, when true values for $\theta$ are unknown, $O$ is used to estimate the $\theta$ that maximizes the likelihood function. That is, $L(\theta|O) = P(O|\theta)$.