# An overview of Statistical Concepts for Machine Learning

Understanding statistical concepts is crucial for machine learning (ML), as many algorithms are grounded in statistics and probability theory. Here's an overview of the **key statistical concepts for machine learning**

### 1. Descriptive Statistics

These summarize and describe the features of a dataset.

- **Mean, Median, Mode**: Measures of central tendency.
- **Variance & Standard Deviation**: Measures of data spread.
- **Skewness & Kurtosis**: Describe data distribution shape.
- **Percentiles & Quartiles**: Describe relative standing in a dataset.

---

### 🎲 2. Probability Theory

Probability is the foundation for most ML models, especially in classification and generative models.

- **Basic Probability**: Events, sample spaces, conditional probability.
- **Bayes' Theorem**: Key for Naive Bayes classifier and Bayesian inference.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- **Probability Distributions**:
  - **Discrete**: Binomial, Poisson.
  - **Continuous**: Normal (Gaussian), Exponential, Uniform.

### 3. Inferential Statistics

These allow us to draw conclusions from a sample about a population.

- **Sampling Techniques**: Random, stratified, etc.
- **Confidence Intervals**: Estimate a range for population parameters.
- **Hypothesis Testing**:
  - Null & alternative hypotheses.
  - Decision making models
  - p-values and significance levels (α).
  - Tests: t-test, z-test, chi-square test, ANOVA.
  - Type I, Type II Error and Power of the test

## 4. Statistical Modeling

Forms the core of many ML algorithms.

- **Linear Regression**: Predicting continuous outputs
- **Multiple Regression and Modeling (various regression models)**
- **Logistic Regression**: For binary classification.
- **Assumptions in Models**: Linearity, independence, homoscedasticity, normality.

**5. Correlation & Causation**

Helps identify relationships between variables.

- **Correlation Coefficient (Pearson, Spearman)**: Measures strength/direction of linear relationships.
- **Covariance**: Measures how two variables vary together.
- **Multicollinearity**: Important in regression (can inflate variances).

**6. Overfitting & Underfitting**

Statistical bias and variance play key roles here.

- **Bias-Variance Tradeoff**:
  - High bias → underfitting.
  - High variance → overfitting.
- **Regularization**: (L1, L2) to prevent overfitting.

---

**7. Evaluation Metrics**

For model performance evaluation, many are rooted in statistics.

- **Classification**: Accuracy, precision, recall, F1 score, ROC-AUC.
- **Regression**: MSE, RMSE, MAE, $R^2$ (coefficient of determination).

---

**Advanced Topics**

- **Resampling Methods**: Cross-validation, bootstrapping.
- **Statistical Learning Theory**: VC dimension, PAC learning.
- **Bayesian Inference**: Update beliefs based on evidence.

- **Time Series Analysis**: ARIMA models, stationarity, autocorrelation.

---

- Examples using Python code?
- A visualization of these concepts?