Cheat Sheet for comprehensive MITx MicroMasters Program in Statistics and Data Science

Data Collection and Sampling

- **Types of Data**
  - **Quantitative**: Numerical values (e.g., age, income)
  - **Qualitative**: Categorical values (e.g., gender, color)
- **Sampling Methods**
  - **Simple Random Sampling**: Each member has an equal chance of being selected.
  - **Stratified Sampling**: Divide population into strata, sample from each.
  - **Cluster Sampling**: Divide population into clusters, sample entire clusters.
  - **Systematic Sampling**: Select every k-th member from the population.

Descriptive Statistics

- **Measures of Central Tendency**
  - **Mean**: Sum of all values divided by the number of values.
- Formula: $\bar{x} = \frac{\sum x_i}{n}$
  - **Median**: Middle value when data is sorted.
  - **Mode**: Most frequently occurring value.
- **Measures of Variability**
  - **Range**: Difference between the maximum and minimum values.
  - **Variance**: Average of the squared differences from the mean.
- Formula: $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$
  - **Standard Deviation**: Square root of the variance.
- Formula: $\sigma = \sqrt{\sigma^2}$

Probability

- **Basic Probability**

- **Probability of an Event**: $P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$

- **Complementary Probability**: $P(A^c) = 1 - P(A)$

- **Conditional Probability**

- **Definition**: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

- **Bayes' Theorem**: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Probability Distributions

- **Discrete Distributions**

- **Binomial Distribution**: $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

- **Poisson Distribution**: $P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$

- **Continuous Distributions**

- **Normal Distribution**: $f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- **Exponential Distribution**: $f(x) = \lambda e^{-\lambda x}$

Inferential Statistics

- **Hypothesis Testing**

- **Steps**:

1. State the null (H_0) and alternative (H_1) hypotheses.
2. Choose the significance level (α).
3. Calculate the test statistic.
4. Determine the p-value.
5. Make a decision (reject H_0 if p-value < α).

- **Confidence Intervals**

- **Mean**: $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

- **Proportion**: $\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Regression Analysis

- **Simple Linear Regression**

- **Model**: $y = \beta_0 + \beta_1 x + \epsilon$

- **Coefficient Estimation**:

- Slope (β_1): $\hat{\beta_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
- Intercept (β_0): $\hat{\beta_0} = \bar{y} - \hat{\beta_1} \bar{x}$

- **Multiple Linear Regression**

- **Model**: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

- **Interpretation**: Each β_i represents the change in y for a one-unit change in x_i holding other variables constant.

Machine Learning

- **Supervised Learning**

- **Classification**: Predict categorical labels (e.g., Logistic Regression, Decision Trees)

- **Regression**: Predict continuous values (e.g., Linear Regression, Random Forest)

- **Unsupervised Learning**

- **Clustering**: Group similar data points (e.g., K-Means, Hierarchical Clustering)

- **Dimensionality Reduction**: Reduce the number of features (e.g., PCA, t-SNE)

Data Visualization

- **Types of Plots**

- **Scatter Plot**: Relationship between two variables.

- **Bar Chart**: Comparison of categorical data.

- **Histogram**: Distribution of numerical data.

- **Box Plot**: Summary of data distribution (min, Q1, median, Q3, max).

- **Tools**

- **Python**: Matplotlib, Seaborn, Plotly

- **R**: ggplot2

- **Tableau**: Interactive dashboards

Data Wrangling

- **Data Cleaning**
  - **Handling Missing Values**:
    - Remove rows/columns with missing data.
    - Impute missing values (mean, median, mode).
  - **Outlier Detection**: Use Z-scores or IQR method.
- **Data Transformation**
  - **Normalization**: Scale data to a fixed range (e.g., Min-Max scaling).
  - **Standardization**: Scale data to have mean 0 and variance 1 (Z-score).

Big Data and Analytics

- **Big Data Technologies**
  - **Hadoop**: Distributed storage and processing.
  - **Spark**: In-memory processing for large datasets.
  - **NoSQL Databases**: MongoDB, Cassandra for unstructured data.
- **Data Pipelines**
  - **ETL**: Extract, Transform, Load.
  - **ELT**: Extract, Load, Transform.

Ethics and Privacy

- **Data Privacy Laws**
  - **GDPR**: General Data Protection Regulation (EU).
  - **CCPA**: California Consumer Privacy Act.
- **Ethical Considerations**
  - **Bias in Algorithms**: Ensure fairness and transparency.
  - **Data Ownership**: Respect user data rights.

Tools and Libraries

- **Python Libraries**
  - **Pandas**: Data manipulation and analysis.

- **NumPy**: Numerical computing.
- **Scikit-learn**: Machine learning algorithms.
- **TensorFlow/PyTorch**: Deep learning frameworks.

- **R Packages**
  - **dplyr**: Data manipulation.
  - **ggplot2**: Data visualization.
  - **caret**: Machine learning.

Practical Tips

- **Data Exploration**

- Always start with descriptive statistics and visualizations.
- Use correlation matrices to identify relationships between variables.

- **Model Evaluation**
  - **Cross-Validation**: Split data into training and validation sets.
  - **Metrics**: Accuracy, Precision, Recall, F1-Score, ROC-AUC.

- **Version Control**

- Use Git for tracking changes in code and data.
- Collaborate effectively with GitHub or GitLab.

Resources

- **Books**

- "The Elements of Statistical Learning" by Hastie, Tibshirani, Friedman.
- "Python for Data Analysis" by Wes McKinney.

- **Online Courses**

- MITx MicroMasters in Statistics and Data Science.
- Coursera, edX for additional specialization courses.

- **Communities**

- Stack Overflow, Kaggle forums for troubleshooting and collaboration.

By Ahmed Baheeg Khorshid

ver 1.0