

Advanced Statistical Methods and Data Mining (Course Outline)

Timothy M. Young
Professor



The University of Tennessee
Center for Renewable Carbon
2506 Jacob Drive
Knoxville, TN 37996-4570
865.946.1119
tmyoung1@utk.edu

Contents

1. **Statistical Process Control (SPC) Review**
2. **Topics in Advanced Statistical Process Control**
 - Autocorrelation – Nonrandom Data
 - Simple Linear relationships
 - Sample correlation coefficient (r)
 - Properties of r
 - Is r statistically significant?
 - **Exercise**
 - Autocorrelation coefficient (lag 1) – r_1
 - Autocorrelation's Effect on Control Limits
 - X-Individual Control Charts Adjusted for Autocorrelation
 - The Exponentially Weighted Moving Average (EWMA) with Control Limits
 - The EWMA statistics and control limits when data are autocorrelated
 - Control Charts for the Coefficient of Variation
 - Multivariate control charts Hotelling's T^2 Statistic
 - **Exercise – Autocorrelation**
 - **Exercise - EWMA**
 - **Exercise – Hotelling's T^2**
 - **Exercise - CV**
3. **Probability Density Functions – “Distributions”**
 - Definitions
 - Discrete Density Functions
 - Binomial Distribution
 - Poisson Distribution
 - Continuous Density Functions
 - Normal Distribution
 - Standard Normal Distribution
 - Chi-Square Distribution
 - T Distribution

4. **Confidence Intervals**

- Introduction
- Confidence Interval on μ when σ^2 is known
- Confidence Interval on μ when σ^2 is unknown
- Confidence Interval on a proportion p

5. **Statistical Significance and Comparative Statistics**

- Hypothesis Testing
- Level of Significance
- Significance Testing
- Significance Tests on the Mean when σ^2 is unknown
- Significance Tests on the Variance
- Comparing Means and Variances (Pooled Tests)
- **Exercise (Density Functions)**
- **Exercise (Confidence Intervals)**
- **Exercise (Comparing Means and Variances)**

6. **Sampling**

- Preliminaries for Sampling
- Sample Design
- Simple Random Sampling
- Sampling Without Replacement
- Sample Size for Estimating Population Mean (μ)
- Sample Size for Estimating a Proportion (p)
- Stratification and Stratified Random Sampling
- Sample Size for estimating the mean for a stratified random sample
- **Exercise – Sample Size for Estimating Sample Mean**
- **Exercise – Stratification and Stratified Random Sampling**

7. **Introduction to Multiple Linear Regression Analysis**

- Simple Linear Regression
- Simple Linear Regression Model
- Estimation of Regression Function Method of Least Squares
- Estimation of Regression Function Least Squares Estimators
- Properties of Fitted Regression Line
- Analysis of Variance (ANOVA) Table for Simple Linear Regression

- Descriptive Measures of Association between X and Y in Regression Model
- Diagnostics for Residuals
- Multiple Linear Regression
- Multiple Linear Regression in Matrix Terms
- Estimation of Regression Coefficients
- Building Regression Models
- Building Regression Models “Criterion for Good Candidates

8. Introduction to Regression Trees

- Theory of Data Mining
- Defining Subspace
- Modeling Subspace
- Theory of the Decision Tree
- Decision Trees
 - Classification Trees
 - Regression Trees
- Decision Tree Modeling Exercises

9. Data Mining and Process Modeling

Data Quality Assessment Techniques

Imputation

Data Fusion

Variable Pre-Selection

Correlation Matrix

Akaike’s Information Criteria (AIC)

Bayesian Information Criteria (BIC)

Genetic Algorithms

Principal Components Analysis

Multicollinearity

Data Mining Methods

Multiple Linear Regression (MLR) Analysis

All Possible Subsets

Regression Trees

Partial Least Squares (PLS)

Neural Networks

Exercises