**Applied Testing for Data Science 2022 Fall**

by Emre Akanak

**Instructor:** Emre Akanak

**Mail:** emre.akanak@tedu.edu.tr

**Phone:** +972587270983

**Office: -**

**Classroom:** Online

**Hour:** 3 Hours

**Code of the Course:** ADS 575

**Name of the Course:** Applied Testing for Data Science

**Content of the Course:**

Testing is one of the main and the essential components of Data Science, which is one of the most frequent applied techniques in daily data science activities. The essential problem is it is only 20% of tests are useful in data science projects and activities and only 13% of Data Product projects ends successfully.

Statical hypothesis testing is a 20th Century methodology adopted by industry and industrial econometric studies but, in reality, has deeper roots than it (the original method was developed in the early 18th Century), and even Karl Pearson's Hypothesis Testing reconstructed it as a Laplacian methodology.

This course focuses applied testing in data science including theoretical background. All tests and testing methodologies will be applied via Python.

**Prerequisites:**

Advanced Calculus, Probability, Distribution, Econometrics Knowledge, Python Codding skills

**Rating of the Course:**

Midterm: 20%

Projects & Presentation 20%

Papers - Projects 40%

Final Exam: 20%

**Week 1: Brief History and Basic Concepts of Testing**

Very brief introduction. Evaluating level of the class for math, statistics, econometrics and Python. Brief discussion about course concepts and knowing the students.

**Week 2: Brief History and Basic Concepts of Testing**

The first (week) lesson starts with historical background of testing concepts (including z-test and t-test conceits). Early applications of hypothesis testing were built by from John Arbuthnot and it was applied by Pierre Simon, Marquis de Laplace. Then in 19th Century it was transformed to modern structure by Karl Pearson.

The lesson also includes “scientific hypothesis” concepts, critiques (Karl Popper’s explanations, critiques against Popperian concept, etc.).

General information on most applied 10 tests by data scientists;

1. Hypothesis Test(ing),
2. A/B Test
3. Chi2 (Chi – Square) Test
4. ANOVA
5. ANCOVA
6. Unit Root Test /Adfuller Test
7. AD (Anderson Darling) Test
8. KS (Kolmogorov Smirnov) Test
9. Mann Whitney U Test
10. Cohen’s – Kappa Test

The lesson does not cover formulations and mathematical concept of tests, it is limited by functions and purposes of the tests. This lesson helps students to have a general idea on tests will help them after coding.

Basic concepts;

Hypothesis Testing (all kinds of tastings) includes structuring Null and Alterative Hypothesis first. The lecture was built on mathematical and statistical fundamentals of testing. Main focuses of the lecture;

Statistical Testing

Software (Excel | Stata | Python)

Python Packages (Statsmodels, SciPy, Scikit-learn)

Basic concepts of computational testing

The basic concepts include;

Confidence Interval

Null Hypothesis and Alternative Hypothesis

Probability Density

Alpha & Betha

Significancy

TP, FP, FN, TN

Type I Error, Type II Error, Type III Error, Type IV Error

Tradeoffs

P-Value

Power

**Week 3: Python**

Basic python coding and basic coding principles. Including;

Variables, and data structures

Functions

Library (as concept)

Basic Libraries as NumPy, Pandas, Scikit-learn, SciPy, Statsmodels

Built-in Functions and User defined functions

Algorithm logic and algorithms concepts

Simple algorithm

Matrix

DataFrame

**Week 4: Distribution**

Statistical tests are based on some assumptions. One of the most basic assumptions is “normal distribution” and this is also one of the most challenging problems which data scientists deals. In nature some phenomenon has normal distribution (height of people, sizes, size of the grains, etc.) but most of social phenomenon have different distribution structures. Such as like news clicks (which has exponential distribution), crime data (some of them has Poisson distribution), etc. Skewness and Kurtosis are as important as standard deviation in order to characterize the distribution.

Data scientists need to understand distribution and, in some cases, apply VST: Variance Stabilization Transformation techniques.

Data scientist should understand the distribution before develop and test hypothesis and see the similarities and differences between categories.

The 4th Week focuses on distributions and Python libraries for understanding the distribution.

**Week 5: Andersen – Darling Testing & Recap of Distribution**

Distribution concept (recap), understanding and visualization the distribution. VST: Variance Stabilization Transformation applications.

Studies of Theodore Wilbur Anderson and Donald Allan Darling

Concept of Andersen – Darling Testing

Implementation

Analysis of Test Results

**Week 5: Hypothesis Testing (Mathematical and Statistical Concept)**

Mathematical and statistical explanation of Hypothesis Testing. Solving examples. Recap of basic concepts and understanding of main logic of Hypothesis testing. Preparation for Python libraries.

The lecture focuses on concept of hypothesis testing, mathematical understanding and computation based on a dataset. The data shows over 80% of hypothesis testing is useless in industry and only 20% of tests provides efficient and realistic results. The focuses of the lecture;

Hypothesis Testing on Python

Null Hypothesis and Alternative Hypothesis Structures

Shaping and creating hypothesis

Conceptional Explanation

Case study on a dataset

Analysis of the case

Bilateral and unilateral hypothesis testing & One Sample and two sample hypothesis testing

**Week 6: Midterm – Exam**

**Week 7: A/B Testing**

A/B Testing is two sample hypothesis testing and it is one of mots used testing techniques in applied data science. The focus of the lecture is applied A/B Test(ing).

Python codes and real-life application of the test.

Analysis of the results

**Week 8: Chi-Square Testing**

Concept of Chai – Square Testing

Implementation

Analysis of Test results

**Week 9: Unit Root Testing**

Unit Testing concept

Unit Root Testing Types

Time Series

Unit Testing basics and understandings

Implementation

Analysis of Test Results

**Week 10: Recap of Testing | Theory | Applications | Problems**

Regarding applicable studies of 6 main testing the lecture focuses of theoretical and applicable recap of the tests and testing concept. The lecture includes;

Recap of testing studies

Reanalysis of results and reevaluation

Brief explanations on other tests, ANCOVA, Mann – Whitney U Test, Cohen’s Kappa Test, etc.

**Week 11: Presentations**

Presentations of projects. (Presentations might require extra hours and additional lessons). If it requires there might be more than one class.

**Week 12: Final Exam**

Regarding the final exam there might be a recap.

**Primary References & Basic Bibliography**

Amemiya,T. (1985)AdvancedEconometrics,Cambridge, MA: Harvard University Press.

Andersen,P.K. and R.D. Gill (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study", The Annals of Statistics, 19, 1100-1120

Andrews, D.W.K. (1990) "Asymptotics for Semiparametric Econometric Models: I. Estimation and Testing",Cowles Foundation Discussion PaperNo. 908R.

Andrews, D.W.K. (1992) "Generic Uniform Convergence", Econometric Theory, 8, 241-257.

Andrews,D.W.K. (1994)"Empirical Process Methods in Econometrics", in: R. Engle and D. McFadden, eds., Handbook of Econometrics, Vol. 4, Amsterdam: North-Holland.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. [Journal Article]. Methodology: European Journal of Research Methods for the Behavioral & Social Sciences, 1(2), 55-70

Bruce P, Bruce A, Gedeck P, Practical Statistics for Data Scientists (50+ Essential Concepts Using R and Python) (2020), O’Reilly, Sebastopol, CA

Downey A, Think Stats, (Exploratory Data Analysis), (2015), O’Reilly, Sebastopol, CA

Fisher, F.M. (1976) The Identification Problem in Econometrics, New York: Krieger.

Fisher, R.A. (1925) "Theory of Statistical Estimation", Proceedings of the Cambridge Philosophical Society, 22, 700-725.

Gruss J, Data Science from Scratch (2019), O’Reilly, Sebastopol, CA

Salsburg D., (2011)The Use of Restricted Significance Tests in Clinical Trials, Springer, Budapest

Taff A., Hypothesis Testing (The Ultimate Beginner’s Guide to Statistical Significance), (2018), TN (Teaching Nerds), New York