Strathmore
UNIVERSITY

iLab Africa

MSc in DATA SCIENCE

COURSE OUTLINE

CODE: DSA 8301 – STATISTICAL INFERENCE IN BIG DATA

Lecturer: Prof. Bernard Omolo **Phone:** +1 864 497 4556

Email: bernardo@strathmore.edu

Consultation: NA **Liaison:** Nicholas Mwadime (0714 542 646)

Purpose: The aim of this module is to give a solid foundation to the understanding of statistics as a general approach to the problem of making valid inferences about relationships using data from observational or experimental studies including classical techniques of hypothesis testing and point and interval estimation, and nonparametric methods.

Intended Learning Outcomes (ILOs):
At the end of this course, the learner should be able to: 1. Describe the principal features of, and differences between, frequentist, likelihood and Bayesian inference 2. Define and derive the likelihood function based on data from a parametric statistical model, and describe its role in various forms of inference. 3. Define, derive and apply different methods for evaluating and comparing estimators 4. Describe, derive and apply lower bounds for the variance of an unbiased estimator 5. Define, derive and apply the error probabilities of a test between two simple hypotheses; define and conduct a likelihood ratio test; state and apply the Neyman-Pearson lemma.
Contact Hours: 45
Prerequisites: DSA 8104, DSA 8202
Lecture time: 17h30 – 20h30 EAT Thursdays

Table 1: Course at a glance.

Content, Outcomes and Activities					
Week / Dates	**Main Topic**	**Intended Learning Outcomes (ILO)**	**Hybrid / Hyflex Learning Activities**		**Assignment and Due Date**
			F2F with TEL	**Virtual**	

Week 1 Mar 20 – 24	Parametric Statistical Models - I	At the end of this topic, students will be able to: - Define statistics as a science - Derive common discrete parametric models - Define the exponential family of probability distributions **Overall ILO: # 2**	N/A	Readings: - Akritas (2015): Ch - 3 - Efron & Hastie (2016): Ch - 5 - Hogg, McKean & Craig (2005): Ch - 3	
Week 2 Mar 27 – 31	Parametric Statistical Models - II	At the end of this topic, students will be able to: - Derive common continuous parametric models - Differentiate Bayesian and frequentist inference **Overall ILO: # 1 and 2**	N/A	Readings: - Akritas (2015): Ch. 3 - Efron & Hastie (2016): Ch - 1, 5	

Week 3 Apr 10 – 14	Fitting Models to Data - I	At the end of this topic, students will be able to: - Define an efficient estimator - Obtain the method of moment estimator of a parameter - Obtain the method of maximum likelihood estimator of a parameter **Overall ILOs: # 2 and 3**	N/A	Readings: - Akritas (2015): Ch – 6	
Week 4 Apr 17 - 21	Fitting Models to Data – II	At the end of this topic, students will be able to: - Derive the least square estimators analytically - Use R to fit a regression model **Overall ILOs: # 2 and 3**	N/A	Readings: - Akritas (2015): Ch – 6	

Week 5 Apr 24 - 28	Evaluating and Comparing Estimators	At the end of this topic, students will be able to: - Obtain the mean square error (MSE) of an estimator - Compare two estimators based on the MSE criterion **Overall ILOs: # 3**	N/A	Readings: - Akritas (2015): Ch - 6	

Week 6 May 1 – 5	Properties of Maximum Likelihood Estimators (MLEs)	At the end of this topic, students will be able to: - Define Fisher information of a parameter - Define the Cramer-Rao lower bound for the variance of an estimator - Define asymptotic efficiency, asymptotic normality and consistency of an MLE **Overall ILOs: # 4**	N/A	Readings: - Hogg, McKean & Craig (2018): Ch - 6 - Akritas (2015): Ch - 6 - Tamhane & Dunlop (2000): Ch - 15 - Hogg, Tanis & Zimmerman (2015): Ch - 6	**Exercise 1**
Week 7 May 8 – 12	Hypothesis Testing - I	At the end of this topic, students will be able to: - Define Type - I and Type - II errors - Perform a simple hypothesis test using the p-value method - Obtain a most powerful test based on the Neyman-Pearson lemma **Overall ILOs: # 5**	N/A	Readings: - Hogg, McKean & Craig (2018): Ch - 8 - Akritas (2015): Ch - 8	Perform a 2-sample hypothesis test for multivariate data using R

Week 8 May 15 – 19	Hypothesis Testing - II	At the end of this topic, students will be able to: - Define a best critical region (BCR) of size alpha - Define a uniformly most powerful (UMP) test of size alpha - Define the likelihood ratio test (LRT) **Overall ILOs: # 5**	N/A	Readings: - Akritas (2015): Ch - 8	

Week 9 May 22 - 26	Chi-square Tests	At the end of this topic, students will be able to: - Derive the multinomial distribution - Understand a goodness-of-fit test based on the chi-square distribution - Perform a goodness-of-fit test for a given data set - Test for independence in contingency tables **Overall ILOs: # 5**	N/A	Readings: - Hogg, Tanis & Zimmerman (2015): Ch - 9	
Week 10 Jun 5 - 9	Interval Estimation	At the end of this topic, students will be able to: - Use the pivot method to construct confidence intervals - Obtain one-sample and two-sample CIs for the mean - Obtain the CIs for regression parameters **Overall ILOs: # 5**	N/A	Readings: - Hogg, McKean & Craig (2018): Ch - 4 - Akritas (2015): Ch - 7 - Hogg, Tanis & Zimmerman (2015): Ch - 7	**CAT – 1** (due 11:59 PM EAT June 11)

Week 11 Jun 12 – 16	Non-parametric Inference	At the end of this topic, students will be able to: - Conduct a sign test for the median - Perform the Wilcoxon rank-sum test for the median - Compare several population distributions using the Kruskal-Wallis Test **Overall ILO: # 5**	N/A	Readings: - Hogg, McKean & Craig (2018): Ch - 10 - Tamhane & Dunlop (2000): Ch - 14 - Hogg, Tanis & Zimmerman (2015): Ch – 8 - Akritas (2015): Ch – 8,9 &10	**Exercise 2**

Week 12 Jun 19 – 23	Bayesian Inference I	At the end of this topic, students will be able to: - Perform Bayesian point estimation - Perform Bayesian interval estimation **Overall ILOs: # 1, 5**	N/A	Readings: - Hogg, McKean & Craig (2018): Ch - 11 - Efron & Hastie (2016): Ch – 3	
Week 13 Jun 26 - 30	Bayesian Inference II	At the end of this topic, students will be able to: - Conduct Bayesian hypothesis testing - Compare two parametric models **Overall ILOs: # 1**		Readings: - Albert, J. (2009): Ch - 8	
Week 14 Jul 3 – 7	Recent Developments in Statistical Inference & Provision for Make-up Classes	At the end of this topic, students will be able to: - Understand k-nearest neighbour (k-NN) and random forests (RF) methods - Apply k-NN and RF to classify multivariate data - Understand the various kernels used in SVM - Perform classification using SVM		Readings: - Efron & Hastie (2016): Ch – 15	**CAT – 2** (due 11:59 PM EAT July 9)

		Overall ILOs: # 1			
Week 15 Jul 10 - 14	Study Week	Study Week		Review for End-Semester Examination	

Course Delivery Methodology

The course will be delivered through a mixture of synchronous lectures via Zoom and asynchronous readings. Synchronous lectures will focus on provision of the conceptual background of inferential techniques and discussion of results from practical and theoretical exercises. Exercises (both theoretical and practical) will be for the student's own practice. Asynchronous readings will be based on specific chapters of the core text books and journal articles and will be undertaken by the students independently. The format will be one 3-hour lecture per week, covering theoretical material.

Academic Assessment

Assessment will comprise Continuous Assessment and End of Semester Examination. Continuous Assessment will comprise two Continuous Assessment Tests (CATs), and exercises to test the students' ability to apply inferential methods to real-life data. The weightswill be distributed as follows: CATs (40%) and End of Semester Examination (60%).

Type	**Weighting (%)**
Examination	**60**
Coursework	**40**
CAT 1	20
CAT 2	20
Total	**100**

Core Reading Materials

1) Buhlmann P. and S. van de Geer, (2011). *Statistics for High-Dimensional Data*; 2011th Edition; Springer; ISBN: 364220191, 1 ISBN: 9783642201912
2) Hastie T., Tibshirani R. and J. Friedman, (2009). *The Elements of Statistical Learning*; 2nd Edition; Springer; ISBN: 0387848576
3) Efron, B., & Hastie, T. (2016). *Computer age statistical inference* (Vol. 5). Cambridge University Press.

4) Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC.
5) Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
6) Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications*. John Wiley & Sons
7) Hogg, R.V., McKean, J.W., & Craig, A.T. (2018). *Introduction to Mathematical Statistics (8th ed.)*. Pearson Education, Boston MA
8) Hogg, R.V., Tanis, E. A., & Zimmerman, D. L. (2015). *Probability and Statistical Inference (9th ed.)*. Pearson Education, Upper Saddle River, New Jersey
9) Akritas, M.G. (2015). *Probability & Statistics with R for Engineers & Scientists (1st. ed.)*. Pearson Education
10) Tamhane, A.C., & Dunlop, D.D. (2000). *Statistics and Data Analysis from Elementary to Intermediate*. Prentice Hall, Upper Saddle River, NJ
11) Albert, J. (2009). *Bayesian Computation with R*. Springer, New York.

Further Reading

1) Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.".
2) Simon, P. (2013). *Too big to ignore: the business case for big data* (Vol. 72). John Wiley & Sons.

Potential Journals

1) *Journal of Big Data*
2) *Big Data & Society*: SAGE Journals
3) *Big Data* | Mary Ann Liebert, Inc
4) *International Journal of Big Data Intelligence (IJBDI)*

Assignments/Exercises

Students will be given exercises during the semester to help them learn the material.

Location and Time

The class will be meeting virtually on Thursdays from 17h30 to 20h30 EAT. Office hours will be held immediately after class or by appointment.

Policies

- **Punctuality** is fundamental. Active **participation** in class discussions is required
- **Plagiarism** is a serious offence. If detected in any form in course work and assignments, the following will apply:
  a. In partial or non-serious cases (such as not citing whole word-for-word quotes), half the total possible marks of the assignment are duly struck off.
  b. In serious cases (such as whole duplication of a paper), a zero policy will apply i.e., all offending assignments will be awarded a mark of zero.

  Note: The level of seriousness referred to above is at the discretion of the lecturer. Appeals are certainly possible through the relevant channels.

Communication Channels

The lecturer can be reached via e-mail, phone, and the Class Liaison (Nicholas Mwadime).

Disclaimer

The instructor reserves the right to make changes to the course outline at any time. These changes will becommunicated to the students in a timely manner.