POS 6933 – Spring 2024: Bayesian Statistics and Data Sciences
Class Number: 16763
Department of Political Science, University of Florida
Thursday: Periods 2-4; TUR 2303

Instructor: Prof. Badredine Arfi
Office: 221 Anderson Hall
Office Hours:
T: 8:30 – 10:30 am; R: 2:30 – 4:00 pm
Or: by appointment thru email

Phone: (352) 273 2357
email: barfi@ufl.edu

COURSE DESCRIPTION AND OBJECTIVES

Imagine that you have a data that you would like to analyze starting with a set of possible explanations. Perhaps, before collecting the data you might have believed that some theories are plausible explanations. After you acquire the data, you decide to examine whether any of those explanations fits the data. Your belief in the credibility of your explanations, given the data, is based on using **Bayes' rule** from probability theory: you update predicted probabilities of an event by using whatever new information comes your way. **This is the essence of doing Bayesian analysis**. And this is what we in fact do every day in reallocating credibility across various possibilities that face us when encountering situations that call for a decision to be made in favor of some of these possibilities – **Bayesian decision is in effect a manifestation of our daily intuitive behavior.** As a matter of fact, you have been doing Bayesian analysis without knowing it. For example, your intuitive interpretation of the usual p-values and confidence intervals are Bayesian, through and through. More generally, the parameters of any statistical model can be estimated using Bayesian methods in a very intuitive way. In sum: two shifts of focus are made in explicitly doing Bayesian data analysis:

(1) We go from a frequentist notion of probability (that is, as a property of the outside world) to a belief-based one (that is, as an observer's belief about observed uncertainty)

(2) We go from point-value hypothesis testing to estimating parameter values and uncertainty by:

   a. Specifying a probability model with some prior knowledge about parameters

   b. Updating our knowledge about the parameters through a conditioning of this probability model on observed data

   c. Assessing the fit of the model to the data and checking the sensitivity of the conclusions to the starting assumptions

Wait: **Bayesian analysis is however not a panacea!** It does not automatically produce THE correct interpretations of the data. Yet Bayesian analysis is built on the observation that all statistical models are subjective – we always make decisi*ons* about variable specifications, significance thresholds, functional forms, error distributions in a nonobjective way, etc.!

The purpose of this course is to introduce and train students to thinking in such a 'Bayesian intuitive' way when doing scientific data analysis. The course starts with the basic concepts of Bayesian analysis and incrementally goes into somewhat more advanced computational methods.

The course takes the perspective of so-called data sciences anchored in Bayesian statistics. Students will therefore be introduced to various methodologies of data sciences known as probabilistic thinking and machine learning. Should time permit we will also briefly discuss neural networks methodologies.

Students are expected to acquire enough skills and understanding of Bayesian statistics such that by the end of the semester they will be able to:

1. Acquire a good understanding of Bayesian methods including Bayesian model specification, Bayesian posterior inference, and model assessment.

2. Use the acquired knowledge of Bayesian statistics to develop and estimate linear and non-linear Bayesian models as well as have enough exposure to MCMC (Markov Chain Monte Carlo) computation.

3. Deploy this knowledge in analyzing data in their respective research fields of interest.

Software packages:

Although social scientists are usually trained in using software packages such as Stata, SPSS, Mplus, SAS, and R, data sciences practitioners find those very limiting and instead use languages such as Python, C++, Julia, and a few others. Therefore, the course will be introducing students to learning Bayesian statistical computations using Python. Not only is Python (like R) a free software, but it does also surpass R by far in the availability of large numbers of powerful packages that make data analysis much richer and more versatile. While surpassing R in sophistication and adaptability, Python is quite flexible in its semantics and comes very close to human natural language in many respects. Moreover, by and large, the field of AI wherever it is applied uses Python, and hence this will equip social scientists to join ranks with data sciences in other fields of knowledge in deploying powerful methodologies of AI to produce theoretical and practical knowledge. More specifically, doing Bayesian analysis using Python packages is quite straightforward and competes with the language Stan (formerly used mostly by statisticians) which R practitioners draw on in developing tools such as BUGS and JAGS to analyze Bayesian models.

The class will be 'walked' into installing and deploying Python and various packages needed for the course in their personal computers during the first week of the semester. In learning how to do Bayesian analysis in any computer language, students have to invest in learning 'software-ways' how to do it, and Python is not unique in this respect.

The purpose and goals of this course is well described in the following quote:

> "Probabilistic programming is a framework that allows you to flexibly build Bayesian statistical models in computer code. Once built, powerful inference algorithms that work independently of the model you formulated can be used to fit your model to data. This combination of flexible model specification and automatic inference provides a powerful tool for the researcher to quickly build, analyze, and iteratively improve novel statistical models. This iterative approach is in stark contrast to the way Bayesian models were fitted to data

> before: previous inference algorithms usually only worked for one specific model. Not only did this require strong mathematical skills to formulate the model and devise an inference scheme, it also considerably slowed down the iterative cycle: change the model, re-derive your inference. Probabilistic programming thus democratizes statistical modeling by considerably lowering the mathematical understanding and time required to successfully build novel models and gain unique insights into your data.
>
> The idea behind probabilistic programming is not new: BUGS, the first of its kind, was first released in 1989. The kinds of model that could be fitted successfully were extremely limited and inference was slow, rendering these first-generation languages not very practical. Today, there are a multitude of probabilistic programming languages that are widely used in academia and at companies such as Google, Microsoft, Amazon, Facebook, and Uber to solve large and complex problems. What has changed? The key factor in lifting probabilistic programming from being a cute toy to the powerful engine that can solve complex large-scale problems is the advent of Hamiltonian Monte Carlo samplers, which are several orders of magnitude more powerful than previous sampling algorithms."[*]

Course strategy:

> "Learn how to think in terms of probabilistic models, and apply Bayes' theorem to derive the logical consequences of our models and data. The approach will also be computational; models will be coded using PyMC, a Python library for Bayesian statistics that hides most of the mathematical details and computations from the user, and ArviZ, a Python package for exploratory analysis of Bayesian models."[*]

Course Objectives/Learning Outcomes:

1. Learn the principles behind Bayesian predictive modeling and model validation to analyze social-scientific data
2. Learn and be able to use Python to apply different classes of statistical and machine learning models
3. Establish command of methods through homeworks and a final research paper
4. Reinforce the use of Python as a statistical computing environment used for Bayesian statistical inference, prediction, scientific computing and data visualization

TEXTS

Hands-on Books using Python Environment:

(Most of these books come with a folder containing codes for all chapters in a jupyter notebook format)

[*] Osvaldo Martin. 2018. Bayesian Analysis with Python Second Edition. Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ. Packt Publishing, forward.

[*] Ibid, preface.

1. Osvaldo Martin. 2018. Bayesian Analysis with Python Second Edition. Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ. Packt Publishing.
2. José Unpingco. 2019. Python for Probability, Statistics, and Machine. Second Edition. Springer.
3. Fabio Nelli. 2018. Python Data Analytics with Pandas, NumPy, and Matplotlib. Second Edition. APress.
4. Oliver Dürr, Beate Sick, and Elvis Murina. 2020. Probabilistic Deep Learning with Python, Keras, and Tensorflow Probability. Manning Shelter Island.
5. François Chollet. 2018. Deep Learning with Python. Manning Shelter Island.
6. Cameron Davidson-Pilon. 2016. Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference. Addison Wesley.
7. Claus Führer, Olivier Verdier, and Jan Erik Solem. 2021. Scientific Computing with Python: High-performance scientific computing with NumPy, SciPy, and Pandas. Second Edition. Packt.
8. Robert Johansson. 2019. Numerical Python: Scientific Computing and Data Science Applications with Numpy, SciPy and Matplotlib. Second Edition. APress.
9. Ashwin Pajankar. 2022. Hands-on Matplotlib Learn Plotting and Visualizations with Python 3. APress.

Bayesian Analysis: Theory

1. Jeff Gill. 2015. Bayesian Methods: A Social and Behavioral Sciences Approach. Third Edition. CRC Press. (On reserve at Library West).
2. Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin. 2013. Bayesian Data Analysis. Third Edition. CRC Press. (On reserve at Library West).

REQUIREMENTS AND ASSESSMENT

The requirement for this course is simple: work diligently and persistently. This includes attending classes, doing the readings carefully before the seminar meets, and working regularly on the computer applications, occasional extra readings, and the research paper. Each student should expect to be spending many hours learning how to excel in using the software packages and use them to estimate models discussed in class.

There will be several homework assignments that the students must complete and upload to canvas. The homework assignments are due on the specified dates; no late submission is acceptable (except with a valid excuse).

A major component of the course evaluation will be a term research paper. Each student will produce a manuscript of high quality using an appropriate modelling strategy.

DISTRIBUTION OF GRADES

1. **35%: Weekly homework exercises**. All assignments are to be uploaded to canvas before the beginning of class on their respective due dates. No late submission will be accepted for any reason (except when justified with university sanctioned documentation). The problem sets will be assigned at the end of the lectures depending on what we cover in the lecture sessions.

2. **15%: Each student will be assigned "presentations**" for the practice session of the course which will consist in presenting the weekly assigned homework (this will be fully explained on the first day of class). A schedule of these presentations will be created on the first day of class.

3. **40%: Research Paper:** Each student is required to choose in consultation with the instructor a research topic. The student is required to find a dataset suitable for the topic and construct a set of research questions. The goal is to produce a high-quality, potentially publishable research manuscript, using a model (or models) discussed in the course, estimated using python packages.

4. **10%: Research Paper Presentation**: Each student will present his/her paper on the last day of classes of the semester. The presentation will consist of a ppt presentation for about ten minutes followed by five minutes of Q & A.

SPECIFICS ON THE RESEARCH PAPER

SPECIFICS ON THE RESEARCH PAPER

For the instructor to provide guidance in the preparation of the paper, students are required to turn in various brief intermediate papers throughout the semester as follows:

1. Find a topic that interests you and a suitable **research dataset** to analyze using a statistical method that falls within the scope of the material covered in this course. Important proviso: you should be guided by the fact that this is a paper for a methods course and hence the emphasis will be put more on the methods part of the paper, and not on the substantive research question that the paper is pursuing (of course, the two aspects are not mutually exclusive). Submit a report summarizing this step of the paper. **Due Date: Jan 25th**
2. Report on the data and various aspects of it. **Due Date: Feb 15th**
3. Begin developing the research design and hypotheses as well as choosing the right statistical model for that purpose. Students are strongly discouraged from using what is commonly called as *logit* or *probit*. These are too simple to provide enough learning challenges, and hence this would defeat the learning goals of the course through a research paper. Submit a short report on this.
   Due Date: March 21st
4. Finalize the research paper focusing mostly on the **methodological aspect** of it without of course neglecting the substantive questions:
   - Show and discuss how you preprocess the data.
   - Construct an appropriate model for the estimation and choose the corresponding python package for that purpose.
   - Carry out a full analysis considering the assumptions and limitations of the model and the data.
   - Draw conclusions on the validity of the model and suggest potential ways to improve your own analysis.

The final paper should be about 15-20 pages long, including the bibliography. **Due Date: April 18th**

5. **Note on the Final Submission of the Paper**:
   Students are required to submit to canvas a **zip** folder that contains the paper (written in a professional format suitable for an academic journal as word, pdf, or latex file), an annotated python Notebook file displaying their complete code and analysis that one would need to replicate the analysis of the paper from beginning to end, and the final dataset used for the paper and, if need be, supplementary materials that are deemed important to understand the paper and its analysis.

The instructor is committed to 'walk the walk' with each student in making his/her research paper a potentially publishable piece.

IMPORTANT DATES

Classes Begin (for this course)	**Thursday, January 11**
Holidays *No classes*	**Monday, January 15: Martin Luther King, Jr.** **March 9-16: Spring Break**
Last Class (for this course)	**Fri, April 18**
ISA Annual Convention – Nashville (no class)	**April 3 - April 6**

	TOPICS –
1	• Introduction: Thinking Probabilistically • Software packages: introduction and installation, etc.
2	• Bayesian Thinking and Programming Probabilistically
3	• Modelling with Linear Regression
4	• Generalized Linear Models
5	• Model Comparison
6	• Mixture Models
7	• Survival Analysis
8	• Causal Inference Analysis
9	• Bayesian Deep Learning
	• Students Presentations

REGULAR IMPORTANT NOTES:

- Incomplete grades may be granted under very special circumstances as supported by valid official documentation (in accordance with the university regulations). Any student seeking such accommodation must request it prior to the deadline for the specific assignment.

- Retroactive extensions/incompletes will not be granted under any circumstances.

- The instructor reserves the right to change any part or aspect of this document should a need for doing so emerge at any point in time during the semester.

- Online course evaluation process: Students are expected to provide professional and respectful feedback on the quality of instruction in this course by completing course evaluations online via GatorEvals. Guidance on how to give feedback in a professional and respectful manner is available from the Gatorevals website. Students will be notified when the evaluation period opens, and can complete evaluations through the email they receive from GatorEvals, in their Canvas course menu under GatorEvals, or via the evaluation system. Summaries of course evaluation results are available to students at the public results website.

- Per university rules there is a zero-percent tolerance on cheating, plagiarism, bribery, misrepresentation, conspiracy, fabrication (see university definitions down below).

- The Writing Studio (352-846-1138) can assist UF students with academic writing through one-on-one consultations either in person or online. Consultations can be scheduled through their website. English language learners can request general writing help or can get help with a specific assignment. are available for students who cannot visit the Writing Studio in person.

UF POLICIES:

- University Policy on Accommodating Students with Disabilities: Students with disabilities requesting accommodations should first register with the UF Disability Resource Center (352.392.8565) by providing appropriate documentation. Once registered, students will receive an accommodation letter which must be presented to the instructor when requesting accommodations. Students with disabilities should follow this procedure as early as possible in the semester.

- Workload: As a Carnegie I, research-intensive university, UF is required by federal law to assign at least 2 hours of work outside of class for every contact hour. Work done in these hours may include reading/viewing assigned material and doing explicitly assigned individual or group work, as well as reviewing notes from class, synthesizing information in advance of exams or papers, and other self-determined study tasks.

- Statement Regarding Course Recording: As in all courses, unauthorized recording and unauthorized sharing of recorded materials is prohibited.

- UF policy on the student computer requirement: Access to and on-going use of a computer is required for all students. The University of Florida expects each student entering a UF Online program, to acquire computer hardware and software appropriate to his or her degree program. Competency in the basic use of a computer is required. Course work will require use of a computer and a broadband connection to the internet, academic advising and registration can be done by computer, official university correspondence is often sent via e-mail and other services are provided that require access through the Internet. While the university offers limited access to computer software through its virtual computer lab and software licensing office, most students will be expected to purchase or lease a computer. The cost of meeting this requirement may be included in financial aid considerations.

- University Policy on Academic Misconduct: Academic honesty and integrity are fundamental values of the University community. Students should be sure that they understand the UF Student Honor Code at http://www.dso.ufl.edu/students.php.

UF STATEMENT ON RECORDING

Students are allowed to record video or audio of class lectures. However, the purposes for which these recordings may be used are strictly controlled. The only allowable purposes are (1) for personal educational use, (2) in connection with a complaint to the university, or (3) as evidence in, or in preparation for, a criminal or civil proceeding. All other purposes are prohibited. Specifically, students may not publish recorded lectures without the written consent of the instructor.

A "class lecture" is an educational presentation intended to inform or teach enrolled students about a particular subject, including any instructor-led discussions that form part of the presentation, and delivered by any instructor hired or appointed by the University, or by a guest instructor, as part of a University of Florida course. A class lecture does not include lab sessions, student presentations, clinical presentations such as patient history, academic exercises involving solely student participation, assessments (quizzes, tests, exams), field trips, private conversations between students in the class or between a student and the faculty or lecturer during a class session.

Publication without permission of the instructor is prohibited. To "publish" means to share, transmit, circulate, distribute, or provide access to a recording, regardless of format or medium, to another person (or persons), including but not limited to another student within the same class section. Additionally, a recording, or transcript of a recording, is considered published if it is posted on or uploaded to, in whole or in part, any media platform, including but not limited to social media, book, magazine, newspaper, leaflet, or third party note/tutoring services. A student who publishes a recording without written consent may be subject to a civil cause of action instituted by a person injured by the publication and/or discipline under UF Regulation 4.040 Student Honor Code and Student Conduct Code.

LEGAL DEFINITIONS

(a) Cheating — The improper taking or tendering of any information or material which shall be used to determine academic credit. Taking of information includes, but is not limited to, copying graded homework assignments from another student; working together with another individual(s) on a take-home test or homework when not specifically permitted by the teacher; looking or attempting to look at another student's paper during an examination; looking or attempting to look at text or notes during an examination when not permitted. Tendering of information includes, but is not limited to, giving your work to another student to be used or copied; giving someone answers to exam questions either when the exam is being given or after having taken an exam; giving or selling a term paper or other written materials to another student; sharing information on a graded assignment.

(b) Plagiarism — The attempt to and/or act of representing the work of another as the product of one's own thought, whether the other's work is published or unpublished, or simply the work of a fellow student. Plagiarism includes, but is not limited to, quoting oral or written materials without citation on an exam, term paper, homework, or other written materials or oral presentations for an academic requirement; submitting a paper which was purchased from a term paper service as your own work; submitting anyone else's paper as your own work.

(c) Bribery — The offering, giving, receiving or soliciting of any materials, items or services of value to gain academic advantage for yourself or another.

(d) Misrepresentation — Any act or omission of information to deceive a teacher for academic advantage. Misrepresentation includes using computer programs generated by another and handing it in as your own work unless expressly allowed by the teacher; lying to a teacher to increase your grade; lying or misrepresenting facts when confronted with an allegation of academic dishonesty.

(e) Conspiracy — The planning or acting with one or more persons to commit any form of academic dishonesty to gain academic advantage for yourself or another.

(f) Fabrication — The use of invented or fabricated information, or the falsification of research or other findings with the intent to deceive for academic or professional advantage.

UF Resources

University Police

The UF police are together for a safe campus. 392-1111 (or 9-1-1 for emergencies) http://www.police.ufl.edu/.

Career Connections Center

Career Connections Center (352-392-1601 | CareerCenterMarketing@ufsa.ufl.edu) connects job seekers with employers and offers guidance to enrich your collegiate experience and prepare you for life after graduation.

Counseling and Wellness Center

Counseling and Wellness Center (352-392-1575) provides counseling and support as well as crisis and wellness services including a variety of workshops throughout the semester (e.g., Yappy Hour, Relaxation and Resilience).

Dean of Students Office

Dean of Students Office (352-392-1261) provides a variety of services to students and families, including Field and Fork (UF's food pantry) and New Student and Family programs

Multicultural and Diversity Affairs

Multicultural and Diversity Affairs (352-294-7850) celebrates and empowers diverse communities and advocates for an inclusive campus.

Office of Student Veteran Services

Office of Student Veteran Services (352-294-2948 | vacounselor@ufl.edu) assists student military veterans with access to benefits.

ONE.UF

ONE.UF is the home of all the student self-service applications, including access to:

- Advising
- Bursar (352-392-0181)
- Financial Aid (352-392-1275)
- Registrar (352-392-1374)

Official Sources of Rules and Regulations

The official source of rules and regulations for UF students is the Undergraduate Catalog and Graduate Catalog. Quick links to other information have also been provided below.

- Student Handbook
- Student Responsibilities, including academic honesty and student conduct code
- e-Learning Supported Services Policies includes links to relevant policies including Acceptable Use, Privacy, and many more
- Accessibility, including the Electronic Information Technology Accessibility Policy and ADA Compliance
- Student Computing Requirements, including minimum and recommended technology requirements and competencies