Calculus Based Introduction to Statistics - **STAT UN 1201**

Section 004

Fall 2019 Syllabus

Lectures: Tuesday, Thursday, 6:10pm- 7:25 pm, **Location:** 602 Hamilton Hall.

Instructor: Samory Kpotufe (skk2175@columbia.edu).

Office hours: Mon. 1:30pm to 3pm. (Location: SSW 911.)
Piazza time: Wed., 5pm to 6pm.

Teaching Assistant: Alessandro Grande (alessandro.grande@columbia.edu)

Office hours: Tues. 1:30pm to 2:30pm, Thurs. 3:30pm to 4:30pm. (Location: SSW 10th floor lounge)
Piazza time: Tues. 2:30pm to 4:30pm, Thurs. 4:30pm to 5:30pm.

Grading Questions: Speak to Alessandro Grande.

1 What should you expect from this course? ℹ

While some Statistics courses focus on how and when to apply various statistical formulas, this course is different: *it's not about formulas, but rather about concepts.* The goal is for you to understand how Statisticians came up with given formulas, i.e., which logical principles they were following. We will use various examples of traditional problems to illustrate concepts, but also more modern problems, e.g., problems from machine learning (computer vision, self-driving cars) that convey general applicability.

- *Our goal:* at the end of the course you should be able to derive your own formulas for new problems.

- *Pace of lectures:* as the name indicates, this course is on the mathematical side, requiring that you keep up every week as the subject *material progresses rather quickly.* Take advantage of lectures, and follow up in office hours to do so.

Text Book

Required: *Jay Devore, Probability and Statistics, 9th Edition.*

2 Prerequisites

A good understanding of calculus, and some multivariate calculus.
Comfort with basic algebraic manipulations (e.g., that $\sum_{i=2}^{9} 2(i-1) = 2\sum_{j=1}^{8} j$ should be obvious).
Basic counting (e.g., remember $n!$, $C(n,k)$ a.k.a $\binom{n}{k}$, $P(n,k)$, and such?) will be assumed.

3 Course Evaluation

Exams (40%) ✍

There will be a midterm and final exam **each worth 20%** of your grade. Any material covered in lectures or in problem sets can make it into the exams. Exams will aim to faithfully test your ability to do homework exercises, and variations thereof, on your own (in fact, exams should be easy if you truly understood homeworks, otherwise you might find exams difficult).

Homework (60%)

Weekly homeworks will be posted Friday evenings and due the following Fridays at 5pm. Each homework consists of a number of problems [1], all worth the same.
Late homeworks will not be accepted, except under extreme circumstances.
Homework questions will be answered on piazza (primary) and office hours.

Collaboration: Homeworks can be discussed (encouraged) but must be written up individually. Homeworks will be *designed* to solidify your understanding of concepts seen in class, and exams will test your ability to do them on your own.

Curving the class

There is a minimal curve: Top 20% will get an A±, Top 50% will get at least a B±. I will decide on lower cutoffs based on overall class performance.
More A's and B's may be given whenever the class performs especially well, e.g., any grade above 90pts/100 is at least an A-, and any grade above 80pts/100 is at least a B-.

4 Overview of topics

We will follow the order of topics below.

Introductory ideas [1 week]

Samples, populations, basic statistics (e.g. sample mean, median) and other ways to summarize data.

Probability Basics [1 week]

Outcome space, Events, and their likelihood, conditioning an event on another, and independent events.

Random variable (r.v.) and important distributions [3 weeks]

Discrete and continuous rvs, mean, variance of a rv, joint distributions, independence.

Principles of Estimation (of distributions)[1.5 weeks]

Unbiased estimation, Maximum Likelihood Principle, Method of Moments.

Confidence Intervals [1.5 weeks]

Intervals under known distributions, and Large sample approaches (Central Limit Theorem).

Hypothesis testing [2 weeks]

Tests under known distributions, and Large sample approaches, p-values and test power.

Regression [2 weeks]

Linear and nonlinear regression, hypothesis testing on function values and confidence bands.

Classification (if time permits) [1 week]

Simple classification rules such as nearest neighbors, perceptron (a simple neural network).

. . .

[1] Some problems will have solutions in the book. This is to encourage you to read the book.