CO245: Probability and Statistics

Din-Houn Lau
Imperial College London
dhl@ic.ac.uk

Autumn 2016

Contents

Course Information

Office Hours

Friday 10–11 in Huxley Room 542 or by appointment

Course Material

All the material for the course will be made available on CATe. Lecture notes will be revealed in batches over the term. Also, Piazza will be available for this course.

Lectures

The lecture notes will be made available ahead of the lectures — you will be expected to print out the notes. During the lecture, I will go through the notes and work through examples and fill in the gaps. You are encourgaged to ask questions during the lecture. Lectures should be available on Panopto – please do not rely on these recordings, as they may fail.

Problem Sheets

There will be non-asssessed problem sheets to complement the material. It is expected that you work through these problems. Solutions will be provided 1 or 2 weeks after the problem sheet is made available.

Problems Classes

One hour a week (Thursdays 1700 - 1800) is a problems class tutorials. In these tutorials you can ask anything about the course material and problem sheets. Myself and GTAs will be present.

Course Assessment

- Exam (11/13): 2 hours, 4 questions. Question 1 is multiple choice.
- Coursework (2/13): tentatively handed out in Week 6 – see CATe.

Course Content

Lecture notes based on slides created by Dr Nick Heard and Dr Nicola Fitz-Simon. All the content (and more) can be found in any introductory statistical textbook such as:

- Grimmett, G. & Stirzaker, D. (2001). *Probability and Random Processes*. Probability and Random Processes. OUP Oxford
- Wasserman, L. (2013). *All of Statistics: A Concise Course in Statistical Inference*. Springer

However all content contained in lecture notes, problem sheets etc.

Chapter 1. Introduction

1.1 Introduction to Uncertainty

This course is about uncertainty, measuring and quantifying uncertainty. Loosely speaking, by uncertainty we mean the condition when results, outcomes, the nearest and remote future are not completely determined; their development depends on a number of factors and just on a pure chance.

Simple examples of uncertainty appear when you bet on the outcome of a football match, turn a wheel of fortune, or toss a coin to make a choice.

Uncertainty appears in virtually all areas of Computer Science and Software Engineering. Installation of software requires uncertain time and often uncertain disk space. A newly released software contains an uncertain number of defects. When a computer program is executed, the amount of required memory may be uncertain. When a job is sent to a printer, it takes uncertain time to print, and there is always a different number of jobs in a queue ahead of it. Electronic components fail at uncertain times, and the order of their failures cannot be predicted exactly. Viruses attack a system at unpredictable times and affect an unpredictable number of files and directories. Uncertainty surrounds us in everyday life, at home, at work, in business, and in leisure.

This course is about measuring and dealing with uncertainty and randomness. It teach you that probability is a language used to describe and quantify uncertainty. But what about Statistics?

1.2 Introduction to Statistics

Definition of "Statistics"

- **Statistics** is the science and practice of developing human knowledge through the use of empirical data.
- **Statistical theory** is a branch of mathematics using probability theory to model randomness and uncertainty in data.
- **Statistical inference** is inference made from the sample data to the defined population using inductive methods and statistical theory.
- A **statistic** is a numerical summary of data.

1.2.1 Population vs. Sample

The previous definitions suggested an important distinction between a sample and a population.

Loosely, we can think of a population as being a large, perhaps infinite, collection of individuals or objects or quantities in which we are interested. For reasons of generality we would wish to make inferences about the entire population.

Example Suppose a new treatment for headaches was being developed by a pharmaceutical company. The target population for this new drug is potentially everybody in the world. To truly, fully understand the efficacy of the new treatment with respect to this population, we would have to administer treatments to every living individual, just after they get a headache, and measure their response. ■

Often it will be impractical or impossible to exhaustively observe every member of a population. So instead we observe what we hope is a representative sample from the population.

To best ensure the sample is representative and not biased in some way, where possible we draw the sample at random from the population.

Example Returning to the drug discovery example, we perform a clinical trial testing efficacy on a small subset of the population, randomising allocation of the new drug or a control treatment. ■

Statistical methods are then used to relate the measurements of the sample to the characteristics of the entire population.

1.3 Probability AND Statistics

A typical Probability problem sounds like this:

Example A folder contains 50 executable files. When a computer virus attacks a system, each file is affected with probability 0.2. Compute the probability that during a virus attack, more than 15 files get affected. ■

Notice that the situation is rather clearly described, in terms of the total number of files and the chance of affecting each file. The only uncertain quantity is the number of affected files, which cannot be predicted for sure.

A typical Statistics problem sounds like this:

Example A folder contains 50 executable files. When a computer virus attacks a system, each file is affected with the same probability p. It has been observed that during a virus attack, 15 files got affected. Estimate p. Is there a strong indication that p is greater than 0.2? ■

This is a practical situation. A user only knows the objectively observed data: the number of files in the folder and the number of files that got affected. Based on that, she needs to estimate p, the proportion of all the files, including the ones in her system and any similar systems. One may provide a point estimator of p, a real number, or may choose to construct a confidence interval of "most probable" values of p.

1.4 Statistical Modelling

Modern statistical methods are largely driven by the notion of a **model**. In contrast with Machine Learning, which focuses on algorithms.

A model is a postulated structure, or an approximation to a structure, which could have led to the data. (Berthold & Hand, 2002)

- Commonly models are **parametric**.
- ⇒ Problem of learning about the underlying population is reduced to learning about a finite set of parameters.

Besides machine learning techniques, computing advances have enabled the fitting of complex parametric/**non-parametric** models (e.g. Bayesian methods, Monte Carlo simulation, . . .).

In this course, we will consider very simple parametric statistical models to represent our populations of interest.

- Statistical inference will thus mean estimating model parameters using our observed sample.
- Likelihood methods will be our main tool for this task. We will learn to calculate the likelihood of a particular parameter solution given our observed sample.