Computing as Experiment
Statistics and Data Analysis

“*Probability is the bane of the age.*”

Anthony Powell
Casanova’s Chinese Restaurant
(from *A dance to the Music of Time*)

Some Etymology

The term "*statistics*" is derived from a corrupted pronunciation of the name

"*de Sade*"

(as in the notorious Marquis).

Experiment in Computer Science

- Many analytic techniques in CS use "*mathematical*" or other *formal techniques* to support claims: eg that *an algorithm satisfies particular performance criteria*; that a *program output is correct with respect to its input*, etc.
- There are, however, a number of activities *not well suited* to such *precise analytic study*.

Justifying that a method "*works well most of the time*"

Fine-tuning system parameters *to boost performance*

Finding support for *machine-learning techniques*.

Addressing these Problems

- In order to deal with questions such as “run-time on average” while, in principle, formal techniques *could* be used sometimes there are a number of *disadvantages*:

  The analysis needed is *extremely complicated*

  as a result it may be *opaque* and *unconvincing*

  it could, even, *be incorrect*

- Using *experiment* can avoid these and

  allow claims to be *presented clearly* (eg by *graphs*, etc)

  so appearing *more convincing* and *accessible*

In this part of the module

- We give an overview of experiment in CS and the basis for this.
- Introduce some simple ideas from statistics by which experimental studies can be analysed.
- These include the notions of

Population and *Sampling*

Expectation, *Variance* and *Standard Deviation*

Estimating Standard Deviation.

Methodology

- Suppose we have some algorithm which “we *suspect*” “*runs efficiently most of the time*”.
- How might we *test* this belief?

1. Run the algorithm a *number of times* on *different data*.
2. Determine the “*average time*” taken.

- This raises several questions

How do we *select* the data?

Data sizes to use?

Number of tests to perform?

How to *present* the *results*?

Population

- The set of items from which objects to test are chosen is called the *population*: usually this is denoted by Ω.
- We limit to *finite size* populations. For example

The set of *permutations* of $\langle 1, 2\ , 3, \ldots, n\rangle$

The set of *directed graphs* with n *nodes*, m *edges*

The set of Year 1 students registered for CS

The set of *US citizens* who are *clinically obese*

The set of *properties* within a *Local Council area*.

Some points to note

- The set forming a *population* may *vary over time* (*students*, *obese Americans*) or remain *fixed* (*permutations*, *graphs*).
- Within a general population we may want to focus on *particular subsets* (eg *female students*, *Band A properties*).
- This may be because the experiment of interest is aimed at gauging characteristics of the *specific* set *relative to the whole* population (eg "*annual income of Band A property owners*")

Sampling

- Ideally, any experiment would involve *every member of the population* being considered.
- This, however, is (usually) *not feasible*.

1. The population may be "*too large*"
2. For "*survey-type*" experiments there may be an *incomplete set of responses* (eg NSS, end-of-module questionnaires)
3. For surveys some responses may be *frivolous*, *untruthful*.

- These force a need to *sample* from a population using a *random selection* approach.

Unbiased vs Biased Sampling

- How do we make selections?
- Each element $X \in \Omega$ has a probability, $\mathrm{P}[X]$ of being chosen.

$$\sum_{X \in \Omega} \mathrm{P}[X] = 1\,;0 \le \mathrm{P}[X] \ \le 1$$

- In an *unbiased* (or *uniform*) selection: $\mathrm{P}[X] = \frac{1}{|\Omega|}$
- Every member has the "*same chance*" of being chosen.
- In *biased* methods some members are *more likely*.

Why use Biased Sampling?

- Despite the fact that "*giving everyone an equal chance*" seems to be "*fair*" the results can be *very misleading* in "*practice*".
- Some algorithms perform *very well* using *uniformly chosen* random graphs but *very badly* on "*graphs in real settings*".
- Some algorithms perform *well* on a *small number* of "*special cases*" but very *badly* on *uniformly chosen random graphs*.
- For example: "*uniformly chosen graphs*" are *dense*; "*real graphs*" are *sparse*.
- "*uniformly chosen binary trees*" are "*deep*"; "*real binary trees*" are "*shallow*".

Random Variables

- Usually we are not so much interested in members of a population in themselves but in the values they report for a particular measure, eg "*the mark obtained by a student*" rather than the *individual student*.
- A *random variable* is just a (we assume Real-valued) function over a population, ie $r : \Omega \to R$.
- Focussing on random variables allows us to define a number of *standard statistical measures*.
- These are the topic of the next lecture.