Statistical Measures
Expectation, Variance, Standard Deviation

Random Variables and Measures

- We recall that a *random variable* is just a (we assume Real-valued) function over a population, ie $r : \Omega \to R$.
- Using a probability distribution, $\mathrm{P} : \Omega \to [0,1]$ the *Expected Value*, $\mathrm{E}[X]$ of a random variable (sometimes referred to as the "*average value*" or "*mean value*") is

$$\mathrm{E}[X] = \sum_{X \in \Omega} \mathrm{P}[X] r(X)$$

- Expectation (relative to the distribution used) is one commonly used notion of "*typical value*": what we *expect* to happen.

Some Examples

- The "expected value" of throwing a *fair die* is $^7/_2$.

$$\mathrm{E}[X] = \sum_{k=1}^{6} \frac{k}{6} = \frac{21}{6} = 3.5$$

- The expected value of throwing a die in which

$$\mathrm{P}[X] = \begin{cases} \dfrac{1}{4} \text{ if } X \in \{1,2,3\} \\ \dfrac{1}{12} \text{ if } X \in \{4,5,6\} \end{cases}$$

$$\mathrm{E}[X] = \frac{6}{4} + \frac{15}{12} = \frac{11}{4} = 2.75$$

- In both cases: $r(X) = X$.

Other Measures – Mode and Median

- The notion of "*average value*" of a random variable (ie expectation) can *sometimes* be *misleading* in assessing behaviour of a population.
- Alternatives are the *mode* and *median*.

Mode of a random variable

$$r(X) : r(X) \text{ occurs } \textit{most often} \text{ in the population}$$

Median of a random variable (finite population)

$$r(X) : \quad r(X) \text{ is the } \textit{middle value} \text{ when ordering}$$

Example 1 – Mode and Median

- Suppose we have Ω as the set of Year 1 students and $r(X)$ is the percentage obtained in an exam.
- If there are 100 students:

10 of whom score 20%

35 of whom score 50%

25 of whom score 60%

30 of whom score 90%

- The *mode* is 50%; *median* mark 60%; *average* 61.5%

Example 2 – Mode and Median

- What if the scores were:

61 score 25%

39 score 100%

- The *mode* and *median* marks are 25%
- but the *average is* 54.25%
- Question: which of these seems “reasonable”?
- *Average* is often used to distort “positive” news: eg “*average earnings*” are significantly higher than “*median earnings*”.
- Question: which of the two is *most often* quoted?

Refinements – Variance

- When studying random variables on a population it is not unusual to find cases where the *average* values are "*similar*" but the *pattern of behaviour* is very *different*.
- eg the examples given on the last slides: *same population*, similar *basis for random variable*, *distinct outcomes*.
- The idea of *variance* (and the related concept of *Standard Deviation*) allows a more careful study of how "average value" is achieved by examining "*how spread out*" the values in a population are.

Variance – Formal Definition

- We have a population, Ω, and random variable, $r(X)$, leading to expectation $\mathrm{E}[X]$ using probability distribution $\mathrm{P}[X]$.
- The *variance*, $Var(X)$, is defined to be

$$\sum_{X \in \Omega} (r(X) - E[X])^2$$

- Variance gives a measure of "*by how much the population as a whole differs from a typical member*".
- Variance is *always non-negative* and the *smaller its value* the *more homogenous* the population is wrt to $r(X)$.

Standard Deviation

- Formally, the *exact Standard Deviation* is: $\sigma \stackrel{\text{def}}{=} \sqrt{Var(X)}$.
- This presents a difficulty in all but very simplified settings.
- Variance is defined wrt to the *whole* population ***BUT***
- It is *not feasible* (or even *possible*) always *to compute* it.
- So we have to "*estimate*".
- In principle we could do this by taking N *samples* from the *population* (according to the *probability distribution*) and compute variance (and standard deviation) using *only these*.
- If we "*take enough samples*" the outcome should be "*close*".

Estimated Standard Deviation

- Take *N samples* from Ω.

$$\langle y_1, y_2, \ldots . y_N \rangle \quad : \quad y_k = r(X_k)$$

- The *estimated Standard Deviation* is:

$$S_N \stackrel{\text{def}}{=} \sqrt{\frac{\sum_{i=1}^{N}(y_i \; - \mathrm{E}[Y])^2}{N}}$$

- Notice that

$$\mathrm{E}[Y] = \frac{\sum_{i=1}^{N} y_i}{N}$$

- ie the expected value of the *sample*.

Informal View

- The idea is to try and capture how the population *overall* behaves by *studying* how a *sample* of its members behave.
- Such approaches are standard in settings such as:

Analysing census statistics

Product quality control

Psephology

- One issue with the form used for "*estimated Standard Deviation*" is that it often (sometimes badly) *underestimates*.
- A device called Bessel's Correction ameliorates this problem.

Bessel's Correction

- Take N *samples* from Ω.

$$\langle y_1, y_2, \ldots. y_N \rangle \quad : \quad y_k = r(X_k)$$

- In *Bessel's Correction for estimated Standard Deviation*:

$$S_N^B \stackrel{\text{def}}{=} \sqrt{\frac{\sum_{i=1}^{N}(y_i - \mathrm{E}[Y])^2}{N-1}}$$

- Standard deviation is the basic tool used to decide "*experimental significance*".

Significance Testing – Informal View

- A typical experiment will have:

A *predicted outcome* (hypothesis): X

An *actual outcome*: Y

- We want to know if "*the chance of our prediction being accurate given the outcome is likely*".
- To assess this:

"*count the number of (estimated) standard deviations by which* Y *differs from* X"

- If "*too many*" the hypothesis is "*not tenable*".

Summary

- Statistical measures and methodology are *important factors* in *CS as an experimental study*.
- Presentation and argument that a given behaviour occurs are derived by *experimental sampling*.
- These are especially needed in fields such as *Machine-Learning* and *Performance Analysis*.
- We also, in addition to raw numerical data, need to find ways to *interpret* this.
- The study of "*Data Fitting*" which we look at next offers some techniques of value.