Statistical Inference

- a measure of confidence

e.g.
How confident are we in our simulation result?

Introduction

by example:
Consider the university student population
Take a random sample of 20 male students and record their height.
Calculate the average for the height.
We call this the sample mean.

e.g $\bar{x} = 175.6$ cm

Suppose someone asks us,
What is the average height of the male student population at UoM?

Our answer will be

$$\mu = 175.6\text{cm}$$

This is a single-value estimate

OR a <u>point estimate</u>

But how accurate are we?

Someone else can take another sample of 20 male student and estimates μ as

$$\mu_0 = 173.2\text{cm}$$

$\therefore$ We might want to give an idea of the precision of the estimate by using an error term

e.g $\mu = 175.6\text{cm} \pm 2\text{cm}$

This means that μ lies somewhere in between

173.6 and 177.6cm

This is called an <u>interval estimate</u>

Nowadays we know that the 'error' or the width of the interval depends on

1) Sample size, n
   Intuitively the larger 'n' is the smaller the error and the smaller the width.

2) The more variable the quantity is
   e.g height
   the larger the standard deviation the greater the error

3) The greater the confidence is the greater the error term

Confidence Intervals

The population mean $= \mu$

μ has a fixed value at a given time

We do not know μ

So we take a sample

Calculate the sample mean $\bar{x}$

Then we specify an Interval within which we are confident that μ lies.

So our answer will be

'We are <u>reasonably</u> confident that μ lies between A and B'

The word <u>reasonably</u> is important.

There is no 100% confidence

100% implies certainty of the value

Even if we specify an interval $[A, B]$ the interval will be so wide that it is useless.

However we can say that 'we are 95% confident of the result'

We have some tools to calculate this.

What does this mean?

On 95% of the occasions when such intervals are calculated, μ, is within the interval specified.

On the other 5% μ will fall outside the interval.

Also if one sample is taken we will not know if the mean is within the interval.

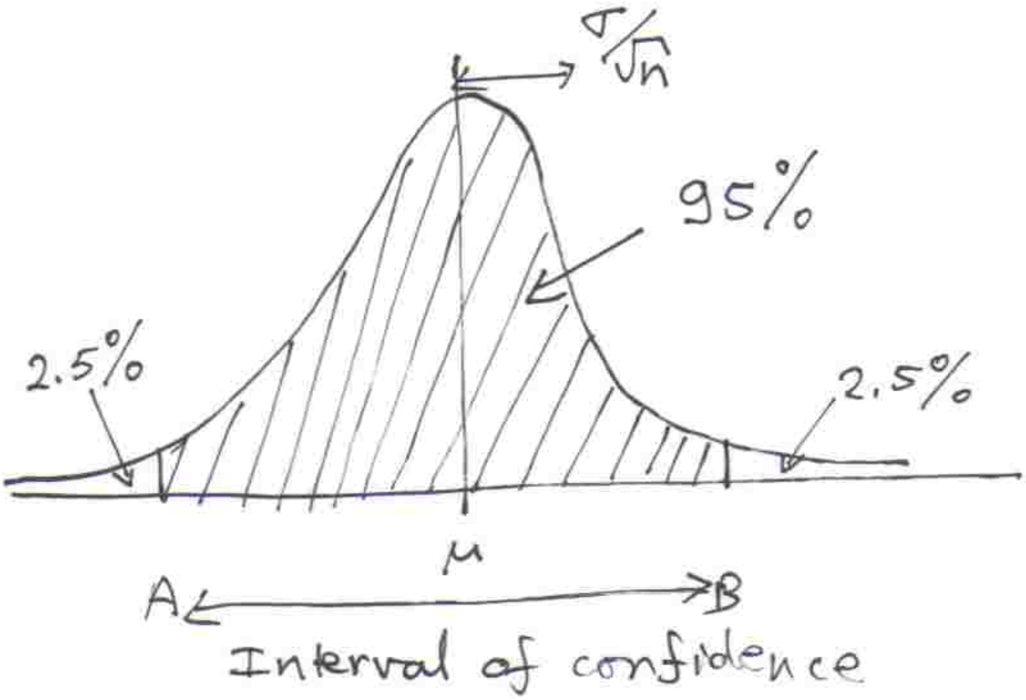
Calculating the Interval

If n is large, say $n > 30$ then the sample mean $\bar{x}$ has a distribution with mean μ and standard deviation of $\frac{\sigma}{\sqrt{n}}$

Also it is approximately normal

[μ and σ pertain to the population]

Below is a graphical representation



From above 95% of the sample mean $\bar{x}$ lies between A and B.

From the tables of the normal distribution this area corresponds to point A and B lying 1.96 standard deviations from the mean of the distribution.

Therefore we state that

95% of the sample means lie in the interval

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \rightarrow \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

OR the probability of μ lying in this interval is 95%

$$P(A < \mu < B) = 0.95$$

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

<u>Now</u> we do not know σ and therefore we need to replace this by s, the sample standard deviation

Then the interval is

$$\bar{x} - 1.96\frac{s}{\sqrt{n}} \rightarrow \bar{x} + 1.96\frac{s}{\sqrt{n}}$$

error term is $\pm 1.96 \frac{s}{\sqrt{n}}$

The reasoning above is <u>acceptable</u> when n is large.

<u>Example</u>

random sample of 30 heights
we calculate $\bar{x} = 175.6$ cm
$s = 8$ cm

then μ lies between

$$175.6 - \frac{1.96 \times 8}{\sqrt{30}} \quad \text{and} \quad 175.6 + \frac{1.96 \times 8}{\sqrt{30}}$$

$$175.6 - 2.86 \qquad 175.6 + 2.86$$

$$172.74 \quad \text{and} \quad 178.46$$

We are 95% confident about this