Sample Distribution

Some Def$^{\underline{n}}$s

Population is the whole set of measurements or quantities we want to sample

Sample is a subset of the population

example

10,000 = UoM Student population

If we are interested in heights then height population = 10,000

If we are interested in male heights then male height population = 4500

A sample is a subset of the above

we can have a random sample or a biased sample

Random sample

Each quantity or measurement in the population has the same probability of being selected.
→ uniform distribution

Biased sample

Probabilities are not the same for each measurement.

Note: When using random sampling we need to devise a method that ensures equal probability of being selected.
e.g. give an identity number to each entity and use a uniform random generator.

Sample Size

This is another important parameter

o 'I am using a sample size of 20'

+ 'that too small, I'm using $n=30$.'

o 'Why?'

+ 'It sounds right, doesn't it!'

Intuitively $n=30$ is more accurate than $n=20$.

But how much accurate?

The answer to this question will be given later....

Sample Distribution

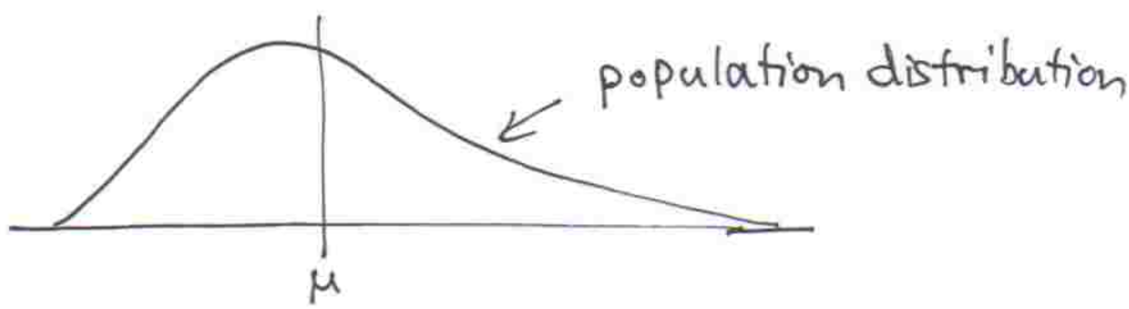
finally ...

Suppose our population has

mean $= \mu$

std dev $= \sigma$

for our measurement, say of height this implies that the distribution is continuous, but not necessarily normal, as below



Suppose we take a random sample of size n.

We then calculate

sample mean $= \bar{x}$

std dev $= s$

We repeat the above experiment.
Then we have ~~another set~~ two more values ~~of~~
for $\bar{x}$ and s

$$\bar{x}_1 \text{ and } s_1$$
$$\bar{x}_2 \text{ and } s_2$$

If we repeat the above for say 40 times
then we have a large number of

$$\text{Sample means} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \bar{x}_{40}\}$$

We can therefore plot a <u>distribution</u>
of the <u>sample means</u>

This distribution will have a mean
and standard deviation

$$\text{mean} = \mu_{\bar{x}}$$
$$\text{std dev} = \sigma_{\bar{x}}$$

Question

What is the connection between
(μ, σ, pdf) for population

and
$(\mu_{\bar{x}}, \sigma_{\bar{x}}, pdf_{\bar{x}})$ for sampling results

?

From statistical theory we have some results.

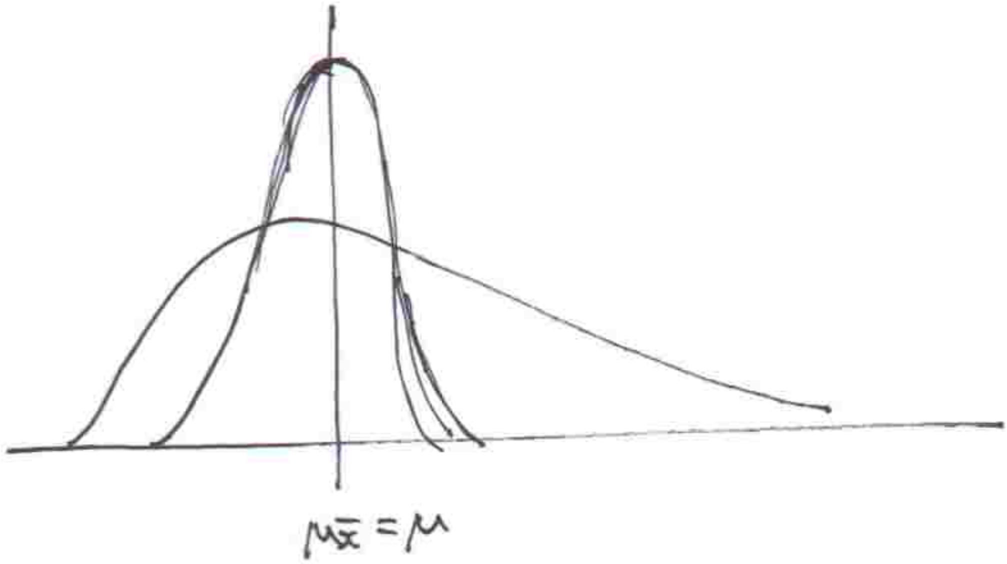
1) $\mu_{\bar{x}} = \mu$

2) $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

3) if n is large, $pdf_{\bar{x}}$ is approximately normal, irrespective of shape of pdf population.

4) if population is normal, then $pdf_{\bar{x}}$ is normal for all n

Graphical view



Result (3) is related to the central limit theorem.

It also emphasises the importance of the normal distribution.