MONASH University

Information Technology

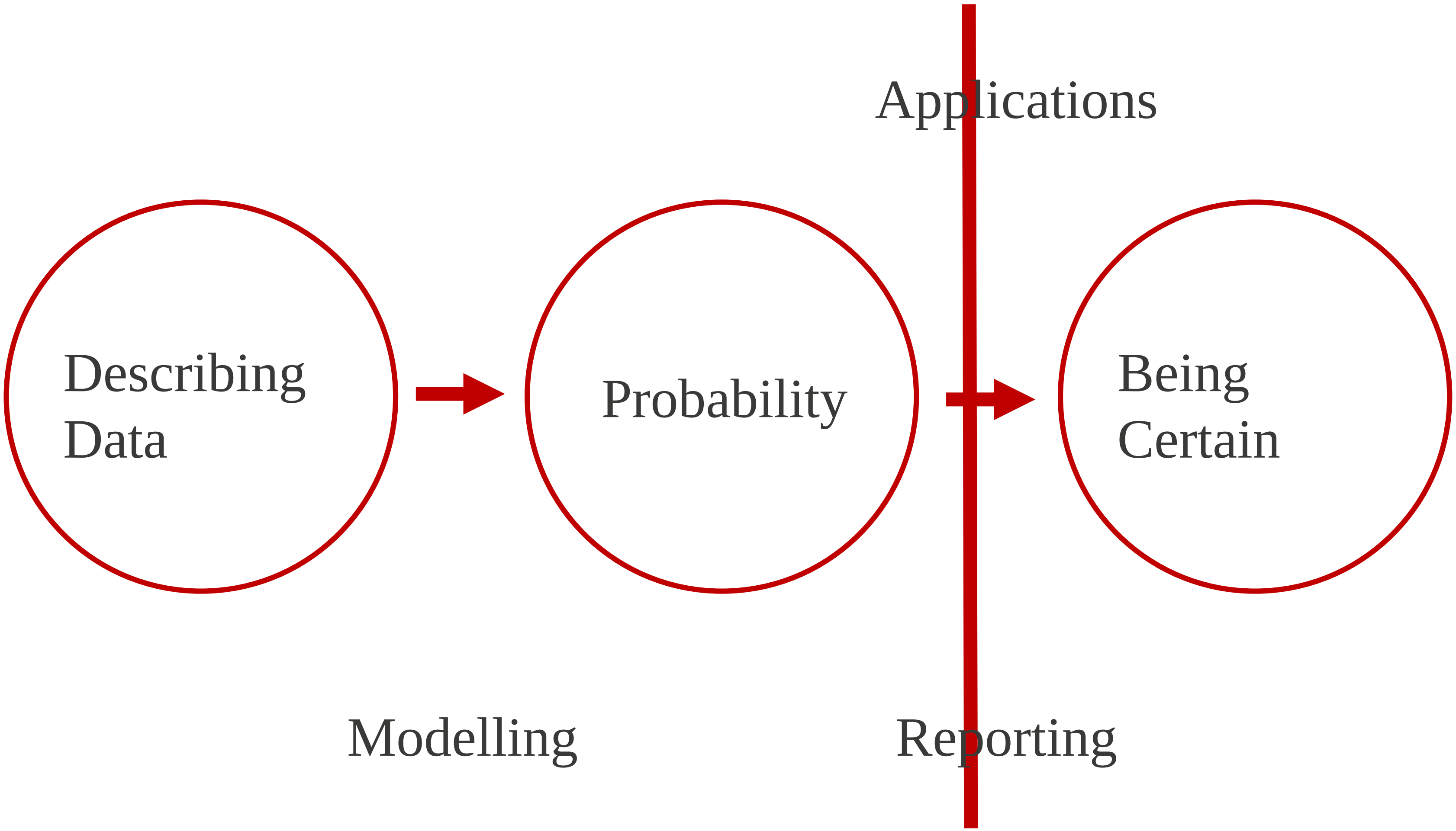
FIT1006
Business Information Analysis

Lecture 14
Theoretical Sampling Distributions

Topics covered:

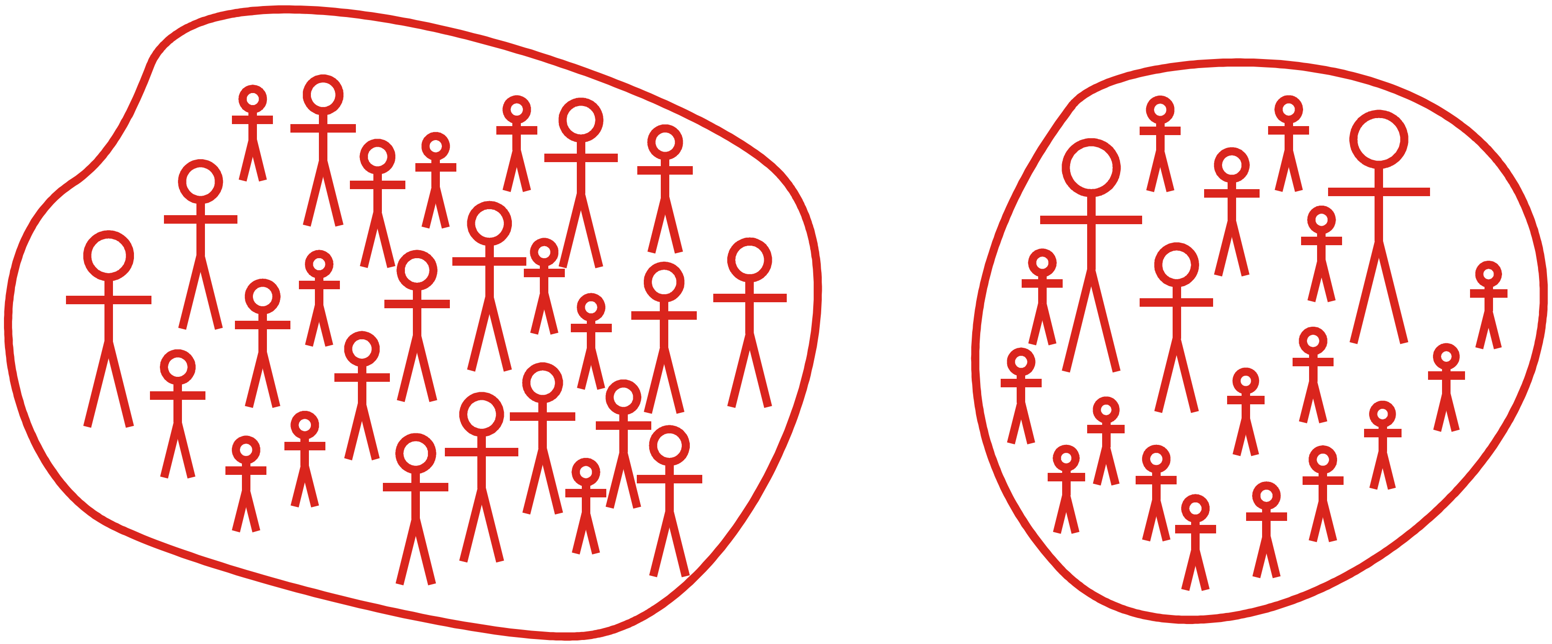
- Theoretical Sampling Distributions
  - Introduction to sampling.
  - The Central Limit Theorem.
  - The sampling distribution of the mean and proportion.

Course outline: Progress report



Update: Being certain

- Two samples are below. Have they come from different populations, or the same? What factors would affect your decision?



Estimating a population parameter

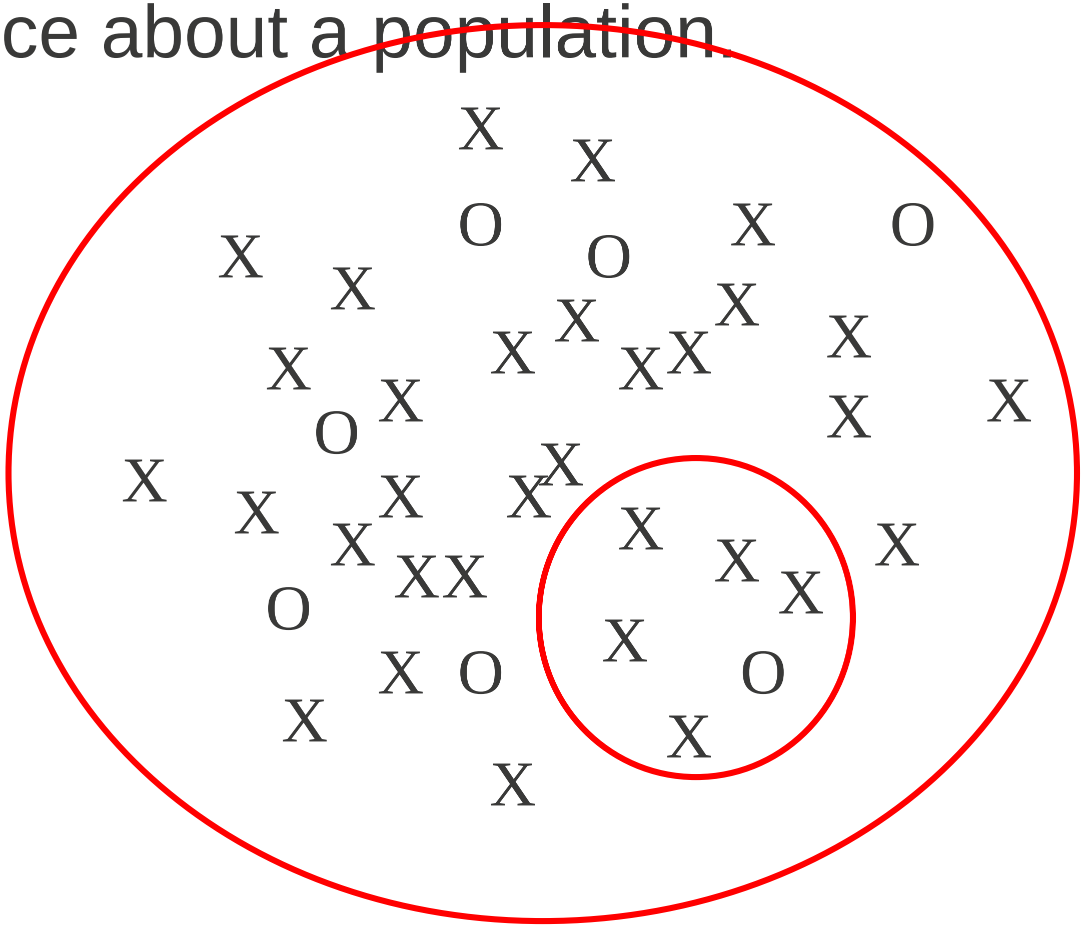
- The usual method of estimating a population parameter is to take a sample, and using the sample statistics make an inference about the population parameter.
- We are frequently interested in the mean of a population, or the proportion of a population exhibiting a certain characteristic.
- We look at how we determine the accuracy of our estimate of these parameters, based on the value of the parameter in question and the sample size.

Estimation

- Part 1. The behaviour of samples

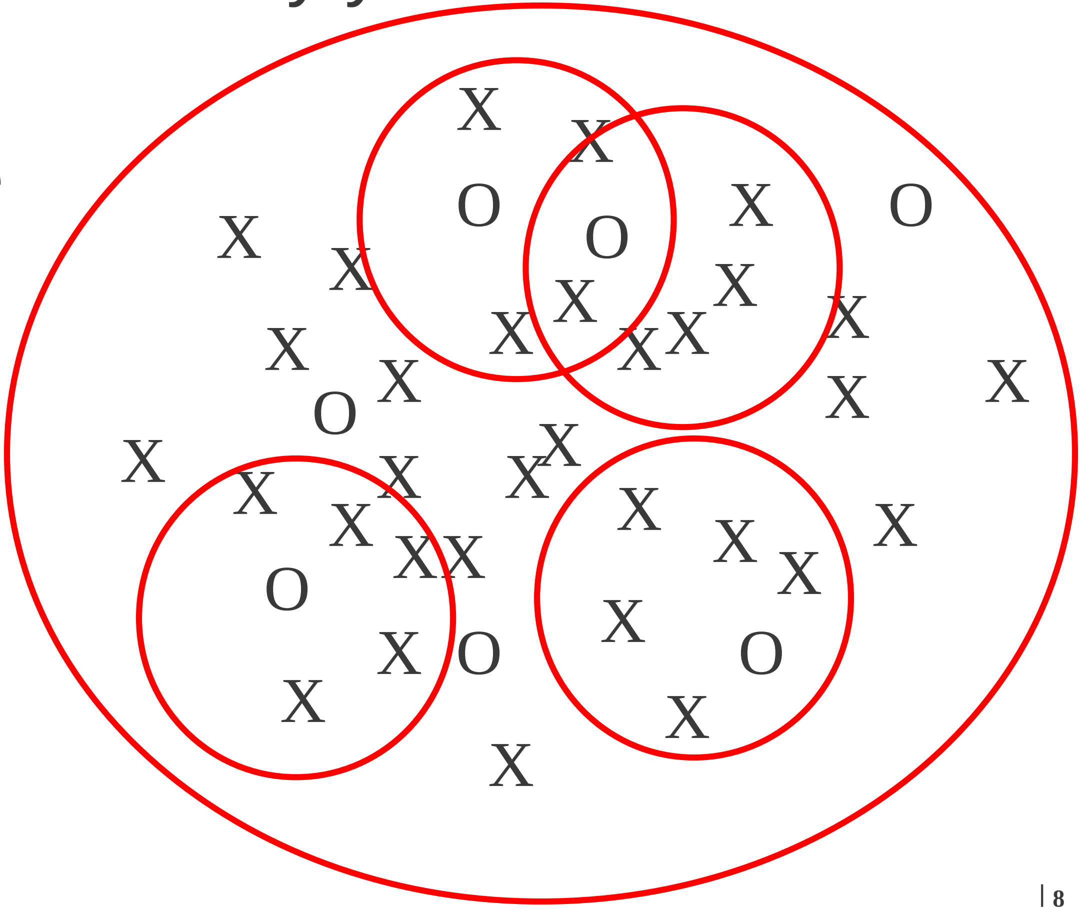
Populations and Samples

- We want to use a sample to make an inference about a population.



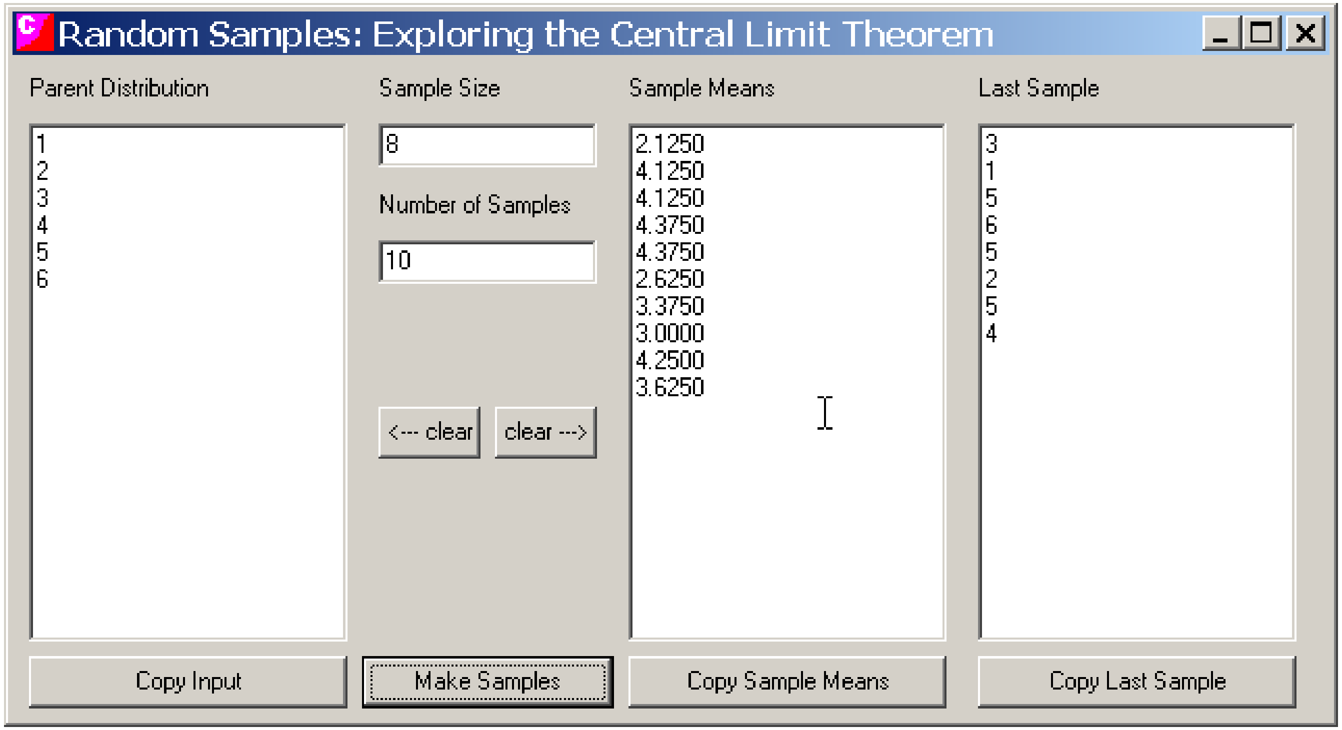
Populations and Samples

- Taking different random samples of the same size from a population may yield different means.
- Thus, the sample mean is itself a random variable having its own distribution.



CLTProject.exe

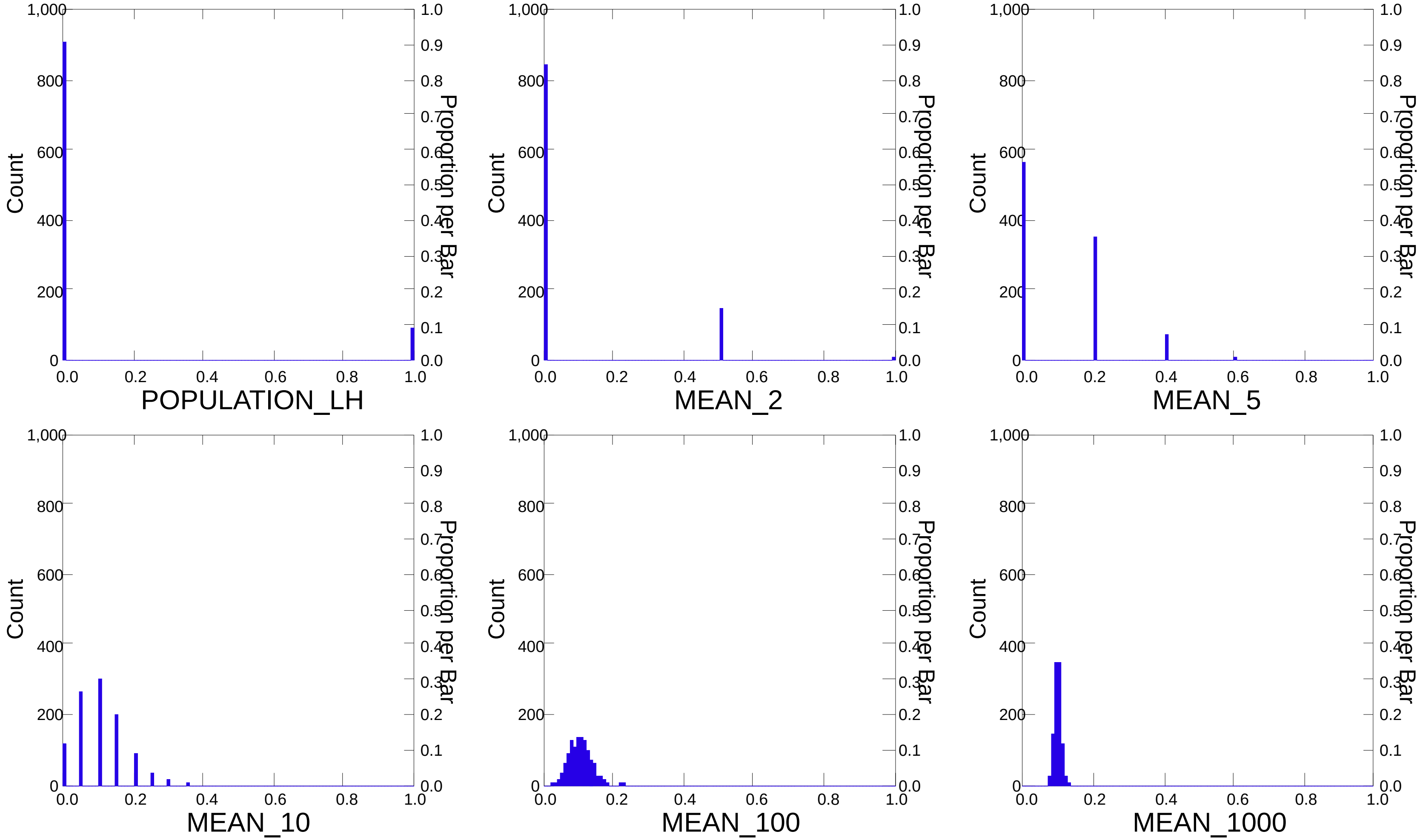
- This application lets you take multiple samples from a population and observe the variability in the samples as a function of sample size. (We will do this in Tutorial 8)



A Binomial distribution problem

- The following slide shows samples taken from a population where, for example:
- 0 = right handed ($p = 0.9$)
- 1 = left handed ($p = 0.1$)
- Samples of size 1, 2, 5, 10, 100, 1000 are taken and the means calculated.
- 1000 samples were taken with replacement. *(That means each sample was chosen observed and put back into the population)*

Effect of sample size



Observations

- As sample size gets larger, 3 things happen:

1. Histogram goes from having a Binomial distribution to approaching a Normal distribution.
2. Sample mean converges to the population mean.
3. Variance of the sample mean decreases – inversely proportional to sample size.

Estimation

- Part 2. The Central Limit Theorem

The Central Limit Theorem

- The Central Limit Theorem is fundamental to inferential statistics.
- The main idea is that if we take large enough sample from a population, we find that *regardless of the distribution of the parent population*, the sample mean is:
- 1. Normally distributed around the population mean.
- 2. The variance of the sample mean is the population variance divided by the size of the sample.

Conditions for the CLT to hold

- 1 Samples must be sufficiently large (n≥30).
- 2 Samples must be of equal size.
- 3 Sampling must be carried out with replacement.

- *In practice we usually only take and analyse one sample from a population. The conditions above are used to establish the validity of the CLT.*

CLT demonstration

- 10000 uniformly [0,1] distributed random numbers were generated using SYSTAT. A histogram of them appears below.

- Data generated using:

  Utilities > Basic >

  BASIC

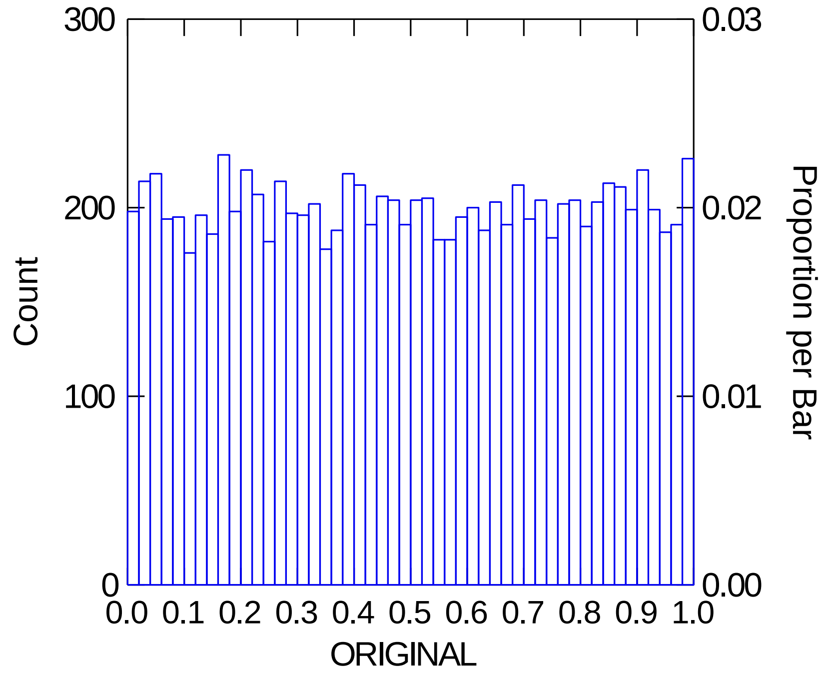
  NEW

  REPEAT=10000

  LET a=URN

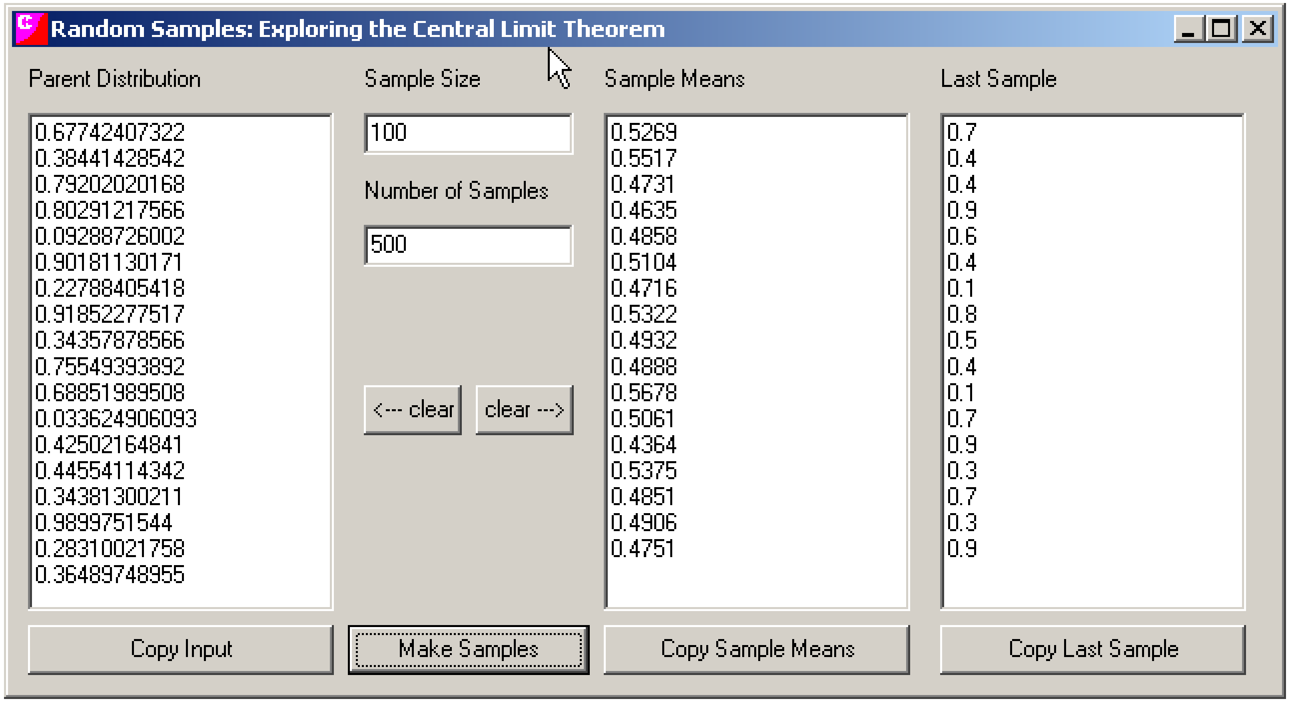
  SAVE d:\Random_10000_Uniform

  RUN



...

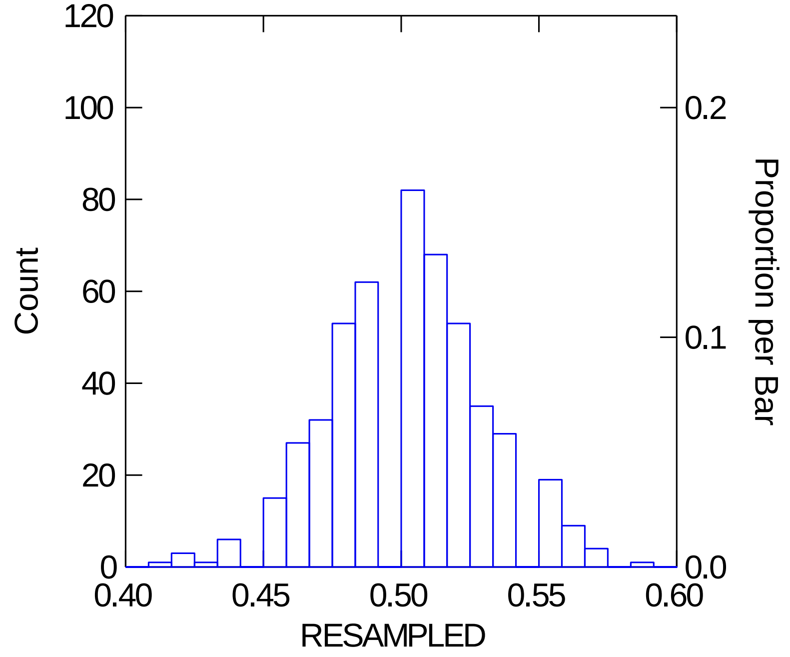
- CLTProject.exe calculates the mean of 500 samples, each of size 100.



- File: FIT1006 Lecture 17 CLT.syz

...

- The randomly generated data was saved as text, copied and pasted into CLTProject.exe. 500 samples of size 100 were taken and the mean calculated.A histogram of the means is below.



...

- Comparing the descriptive statistics for both the original data and the 500 samples of size 100.

	ORIGINAL	RESAMPLED
N of cases	10000	500
Minimum	0.000	0.410
Maximum	1.000	0.590
Median	0.499	0.500
Mean	0.501	0.501*
Standard Dev	0.290	0.028*
N 1 of 4	0.247	0.480
N 2 of 4	0.499	0.500
N 3 of 4	0.754	0.520

Estimation

- Part 3. The sampling distribution of means and proportions

Notation, main characters:

Parameter	Population	Sample
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	π	p

$\sigma_{\bar{x}}$ = standard error of the sample mean

σ_p = standard error of the sample proportion

The sample values are used to estimate the <u>unknown</u> population parameters, taking into account variability introduced by sampling.

The Sampling Distribution of the Mean

From the CLT, if we take a sample of size n,

From a population with mean μ and variance σ^2

Then, as n increases:

The sample mean, $\bar{x} \rightarrow \mu$, and variance$(\bar{x}) \rightarrow \dfrac{\sigma^2}{n}$

thus $\bar{x} = \mu$ and $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ for n large.

sample standard dev.
(standard error)

https://flux.qa (Feed code: SJ6KGV)

Question 1

If a sample of 100 accounts is taken from a population, with mean = $2000 and standard deviation $500; the distribution of the sample mean is:

A. Normal(mean = 20, stdev = 5)

B. Normal(mean = 20, stdev = 50)

C. Normal(mean = 2000, stdev = 5)

✓ D. Normal(mean = 2000, stdev = 50)

Example 1

- A sample of 100 accounts were taken from a population of accounts with mean = \$2000 and standard deviation \$500. What is the probability that the sample mean will be less than 2050?

(Z<z) for Z =

z	0.00	0.01	
0.0	0.5000	0.5040	0.50
0.1	0.5398	0.5438	0.547
0.2	0.5793	0.5832	0.5871
0.3	0.6179	0.6217	0.6255
0.4	0.6554	0.6591	0.6628
0.5	0.6915	0.6950	0.6985
0.6	0.7257	0.7291	0.7324
0.7	0.7580	0.7611	0.7642
0.8	0.7881	0.7910	0.7939
0.9	0.8159	0.8186	0.8212
1.0	0.8413	0.8438	0.84
1.1	0.8643	0.8665	0.8
1.2	0.8849	0.8869	
[illegible]	0.9032	0.904	

$$\text{From the population, } \mu = 2000, \ \ \sigma = 500$$

$$\text{For the sample, } \bar{x} = 2000, \ \ n = 100$$

$$\text{thus } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{100}} = 50$$

$$\text{and } \bar{x} \approx N(2000, 50^2)$$

$$P(\bar{x} < 2050) = P\left(z < \frac{2050 - 2000}{50} \right)$$

$$= P(z < 1), \ \ z \approx N(0, 1^2) = 0.8413$$

The Sampling Distribution of a Proportion

If we take a sample of size n,

From a population with proportion p of interest

Then, from the CLT, as n increases:

Sample proportion, $p \rightarrow p$, variance(p) $\rightarrow \dfrac{p(1-p)}{n}$

Thus $p = p$, $s_p = \sqrt{\dfrac{p(1-p)}{n}}$ for n large,

$np,\ n(1\text{-}p) \geq 5$

standard error of proportion

Example 2

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486

- It is thought that the proportion of left handed people in the population is 10%. What is probability that a sample of 100 people taken at random would have a proportion of left handers less than 0.12?

$$\boxed{\frac{\pi(1-\pi)}{n}}$$

$$\pi = 0.1,\ n = 100$$

$$E(p) = 0.1,\quad Var(p) = \frac{0.1 \times 0.9}{100} = 0.03^2$$

$$\text{thus } p \approx N(0.1, 0.03^2)$$

$$P(p < 0.12) = P\left(z < \frac{0.12 - 0.1}{0.03}\right)$$

$$= P(z < 0.67),\ z \approx N(0, 1^2) = 0.7486$$

Reading/Questions (Selvanathan)

- Sampling inference and sampling distributions.
  - Reading: 7th Ed. Chapter 9.
  - Questions: 7th Ed. 9.4, 9.12, 9.13, 9.18, 9.24, 9.25