Session 2

Sampling distributions and central limit theorem

ISB

Managerial Decisions

What is the amount of time spent by our potential customers on the web?

What is success rate and demand for the drug?

What are the number of man hours required to complete such a project?

How many new customers will I acquire if I open a store in this area?

What is the impact of a stock-out on consumer behavior?

Will our quality improve after the consulting engagement?

Will our order lead times gone down after the merger?

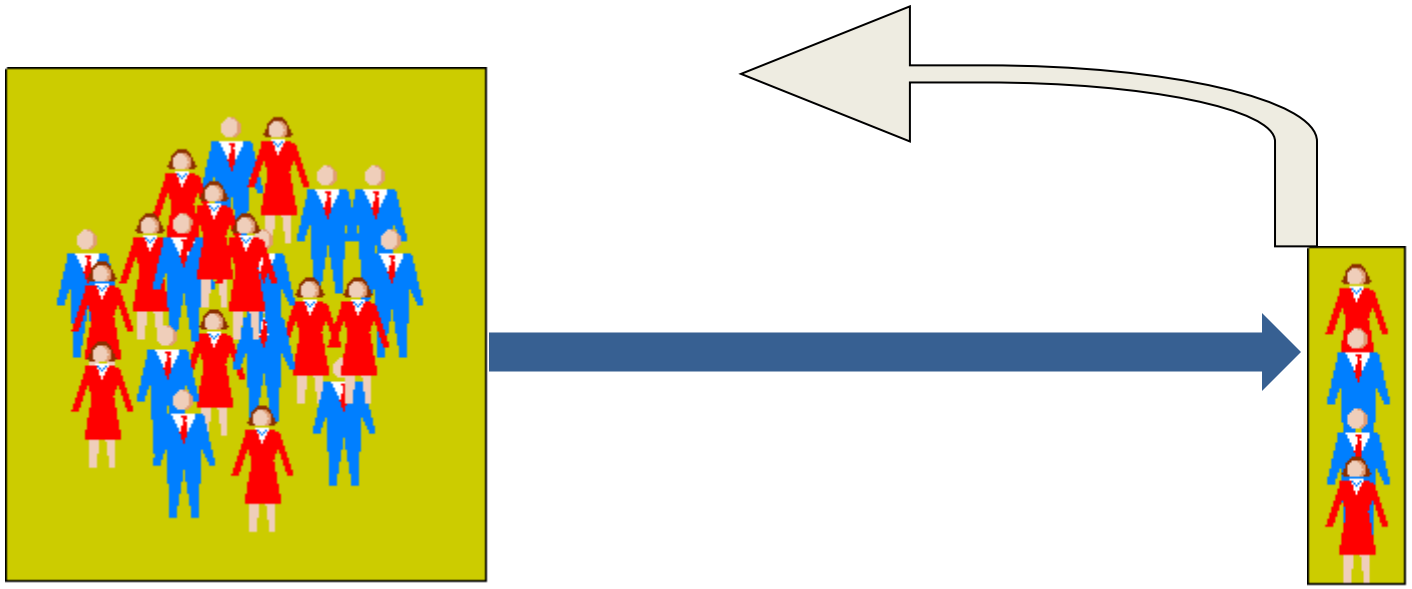
ISB

Learning objectives

- What is statistical inference?
- How to (and how not to) choose a sample?
- What are sample statistics and their properties?
- What is the central limit theorem and how is it useful?

Statistical Inference

Statistical Inference: Make statements about the characteristic of population on the basis of a sample

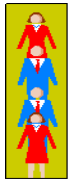


Population
Total collection of objects or people to be studied (or set of all information of interest to the decision maker)

Sample
Subset of a population

Population parameters and Sample statistics

- **Population parameter**: Characteristic of the population
- **Sample statistic:** Characteristic of the sample



Population Parameter		Sample Statistic
μ	Mean	$\bar{x}$
σ^2	Variance	s^2
π	Proportion	p

Some sample statistics might be used as a point estimate for a population parameter

How to select the sample

Example: What is the average work experience of all entering PGP students in this batch?

- Sample should be representative of the population.

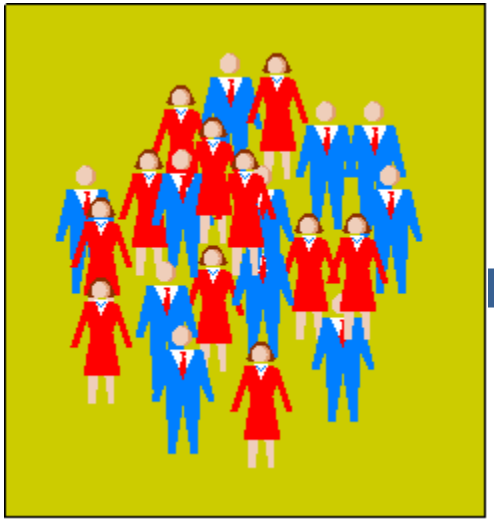
In a Simple Random Sample (SRS) each unit has an equal chance of being in the sample and each unit is selected independently

- How to obtain an SRS? Prepare a list and use a randomization device (like the RANDBETWEEN function in excel)

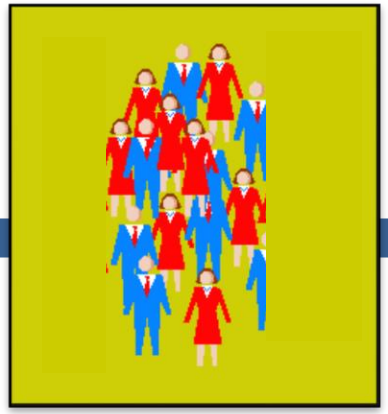
ISB

Sampling frame

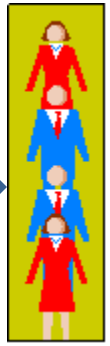
Statistical Inference: Make statements about the characteristic of population on the basis of a sample



Population
Total collection of objects or people to be studied (or set of all information of interest to the decision maker)



Sampling frame
Collection of objects or people from the population who can be sampled



Sample
Subset of a population

Selecting the sampling frame

- Does the sampling frame represent the population?
  - e.g. Literary Digest vs. George Gallup polls
- The available list may differ from desired list
  - e.g. we don't have list of customers who did not buy from a store, list of hourly wages of all convenience store employees in the city
- Sometimes, no comprehensive sampling frame exists
  - e.g. when forecasting for the future, individuals who are reluctant to be identified as having a particular characteristic.

ISB

Typical Pitfalls in Sampling

- Collecting data only from volunteers (voluntary response sample)
  - e.g. online reviews (yelp.com, maps.google.com, tripadvisor.com)
- Picking easily available respondents (convenience sample)
  - e.g. choosing to survey in In-Orbit mall
- A high rate of non-response (more than 70%)
  - e.g. CEO / CIO surveys on some industry trends

ISB

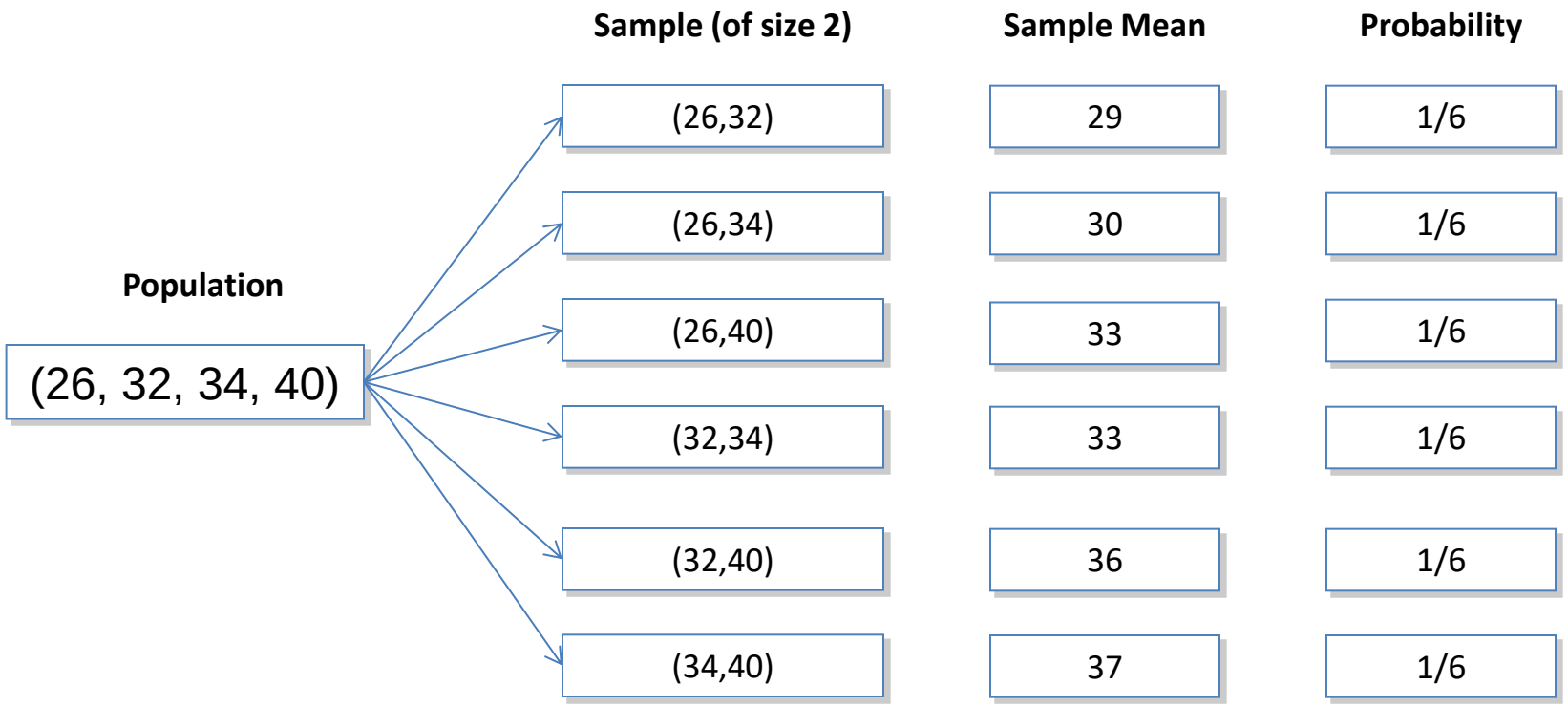
Sampling variation

- What is the average work experience of all participants of the BA course?

	Sample 1	Sample 2
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
Sample mean		

Sampling variation

- Sample mean varies from one sample to another
- Sample mean is a random variable
- Sample mean can be (and most likely is) different from the population mean



Central Limit Theorem (CLT) & the distribution of the sample mean

- The distribution of the sample mean
  - will be normal when the distribution of data in the population is normal
  - will be approximately normal even if the distribution of data in the population is not normal, under some conditions

- Mean $(\bar{X}) = \mu$ (the same as the population mean of the raw data)

- Standard deviation $(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation and n is the sample size
  - This is also referred to as Standard Error of the Mean and is also denoted by $SE(\bar{X})$ or $\sigma_{\bar{X}}$

CLT is Valid When…

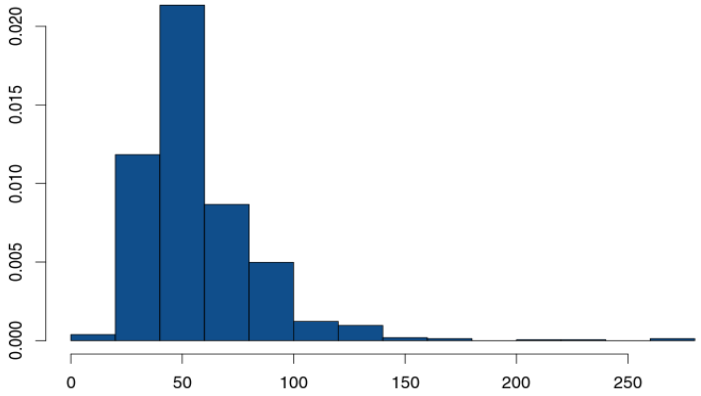
- Each data point in the sample is independent of the other
- The sample size is large enough
- A sample size of 30 is usually considered large enough but there are more precise conditions
  - $n > 10\ (K_3)^2$, where K_3 is sample skewness, and
  - $n > 10\ |K_4|$, where K_4 is sample kurtosis
- Adequate sample size depends on the distribution of data – primarily its symmetry and presence of outliers
- If data is quite symmetric and has few outliers, even smaller samples are fine. Otherwise, we need larger samples

ISB

Be careful…Tale of two distributions

There are two distributions:

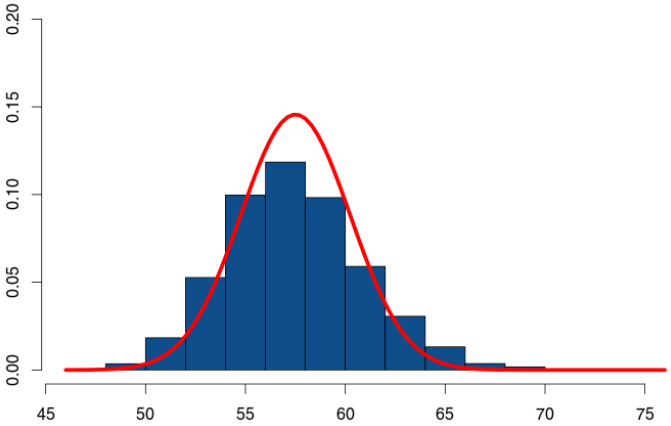
1. Distribution of population data (e.g. the work experience data of the students of this batch)
   - Indicates the variation in population data



Mean (μ) = 57.50
Std. Dev. (σ) = 27.39

Sampling distribution

2. Distribution of sample means across all possible samples (e.g. the average work experience data of samples of size = 64 students)
   - Indicates how uncertain we are about $\bar{X}$



Mean = 57.46 ≈ μ
Std. Dev = 3.42 ≈ σ/√64

Summary of Session II

- What is statistical inference?
- Statistical inference is the process of making probabilistic inferences about population parameters based on sample statistics

- How to (and how not to) choose a sample?
- You want a simple random sample. To do so you require a sampling frame that represent the population and a randomization device

- What are sample statistics and their properties?
- Sample statistics are random variables because they vary across samples drawn from the same population. They can be used as point estimates of the population parameter

- What is the central limit theorem and how is it useful?
- Central limit theorem implies that no matter what the population distribution is, the sample mean ($\overline{X}$) is normally distributed with mean (μ) and standard error $\left(\frac{\sigma}{\sqrt{n}}\right)$, approximately.

ISB