Inferential statistics

Shyue Ping Ong

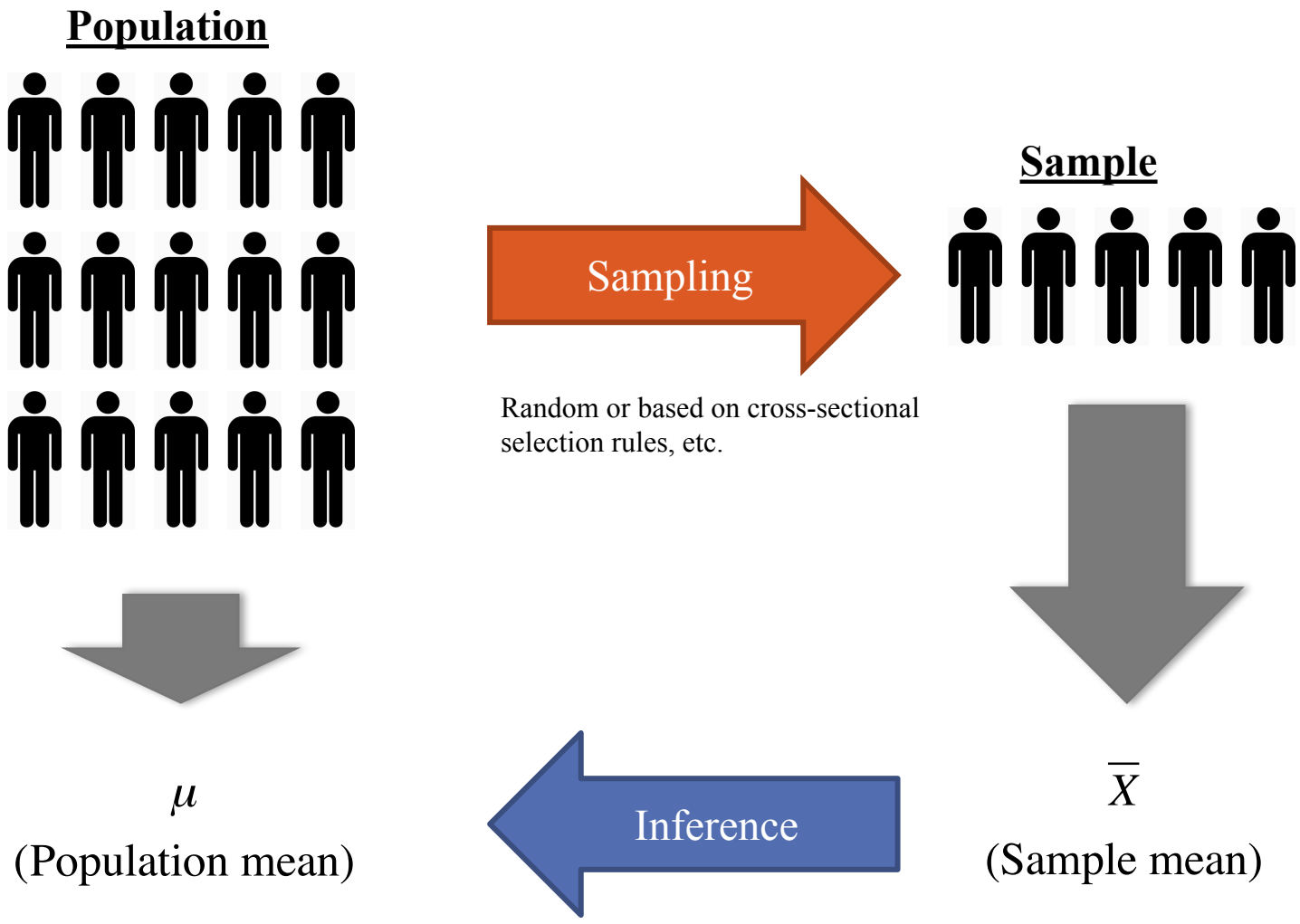
UC San Diego
Jacobs School of Engineering

Readings

Chapter 5

Key concept in inferential statistics

E.g., all voters, all potential samples of a material, etc.



Populations

Complete set of observations (or potential observations)

Examples

- GPA scores of all currently enrolled students at UCSD
- Presidential preferences of all currently registered voters
- US Census
- Yield strength of all possible samples from a block of steelμ

Samples

Subset of observations from a population

Typically much smaller than the population

Examples

- Bureau of Labor Statistics Monthly Unemployment Survey only uses 1% of US population
- Gallup polls of presidential elections typically uses ~4000 voters

Random sampling

For valid application of inferential statistical techniques, samples must be random

Random implies that at each stage of sampling, the selection process guarantees that all potential observations in the population have an equal chance of being included in the sample

Random assignment of subjects

Typically used in medical experiments involving subjects

Though subjects cannot be selected randomly from real population, they can be assigned randomly, with equal likelihood, to a treatment and a control group

Sampling Distribution

Sampling distribution of the mean is the probability distribution of means of all possible random samples of a given size from some population

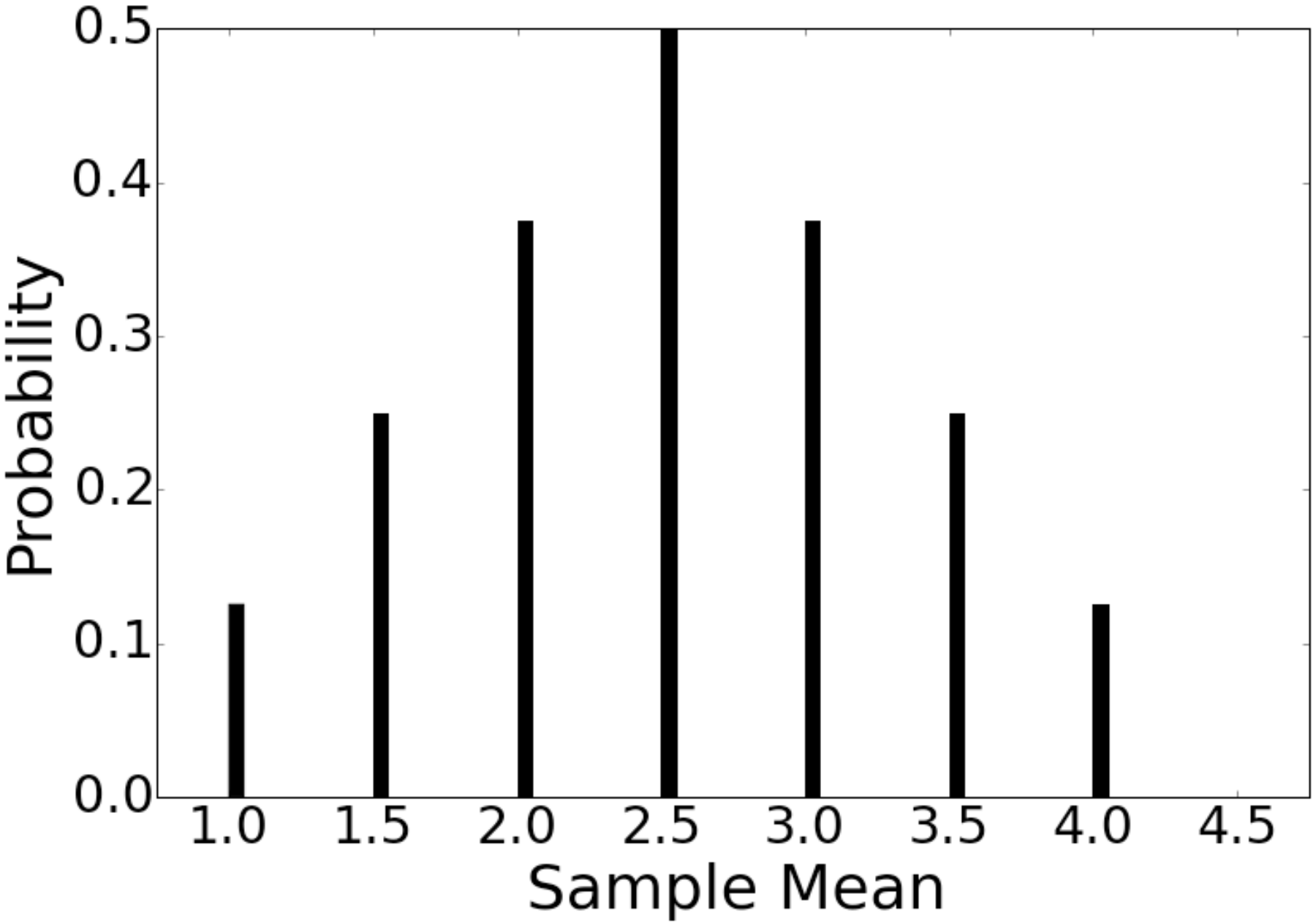
The *most important* concept in inferential statistics

Example

What are the means of all possible samples of a roll of two four-sided dice?

i. How many possible samples are there?

First dice	Second dice	Mean
1	1	1
1	2	1.5
1	3	2
1	4	2.5
2	1	1.5
2	2	2
2	3	2.5
2	4	3
3	1	2
3	2	2.5
3	3	3
3	4	3.5
4	1	2.5
4	2	3
4	3	3.5
4	4	4


0.5
0.4
0.3
0.2
0.1
0.0
Probability
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Sample Mean

Notation

$\overline{X}$: Sample mean

s: Sample standard deviation

μ : Population mean

σ: Population standard deviation

$\mu_{\overline{X}}$: Mean of all sample means

$\sigma_{\overline{X}}$: Standard deviation of all sample means

Mean and standard deviation of all sample means

The mean of all sample means is always equal to the population mean.

$$\mu_{\bar{X}} = \mu$$

The standard deviation of all sample means is given by:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

This is a measure of how much the sample means deviate from the population mean. **As sample size increases, the standard deviation of the sample mean decreases.**

Central Limit Theorem stated

$$\text{For large n, } \overline{X} \sim N(\mu, \sigma_{\overline{X}}^2) \sim N(\mu, \frac{\sigma^2}{n})$$

If X is itself normally distributed, the approximation is exact for all values of n, i.e., the sample mean has a normal distribution regardless of n. (Note that the std deviation is still large for small n.)

If X is non-normal, typically a sample size of 25-100 is sufficiently large to ensure that the normal approximation holds.

Estimating parameters

In many science and engineering problems, you are interested in estimating an unknown population mean from a series of experiments or samples.

- E.g., estimating the true population yield strength of a material by taking samples from a block and conducting tests
- Estimating the real reaction rate but conducting several measurements of the reaction rate under the exact same conditions.

Point estimation

Using a single value to represent the unknown population mean.

From the central limit theorem, we know

$$\text{For large n, } \overline{X} \sim N(\mu, \sigma_{\overline{X}}{}^2) \sim N(\mu, \frac{\sigma^2}{n})$$

We can therefore use $\overline{X}$ as an estimate for the true population mean. However, this provides no information about the degree of inaccuracy of this estimate due to sampling variability.