Introduction to Statistical Modelling

STAT2507D

Chapter 7-I

Sampling Distributions

INTRODUCTION

- Parameters: numerical descriptive measures for populations
  - For the normal distribution, the location and shape are described by μ and σ
  - For binomial distribution consisting of n trials, the location and shape are determined by p
  - In practical situations, which type of probability distribution (binomial, normal or others) can be used as a model, is known

EXAMPLES

- Example 1: A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown
- Example 2: An agronomist believes that the yield per hectare of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown

INTRODUCTION

- Often the values of parameters that specify the exact form of a distribution are unknown
- We must rely on the sample to estimate these parameters
- This sample must be selected in a certain way so that this sample provide reliable information.

Sampling Plans and Experimental Designs

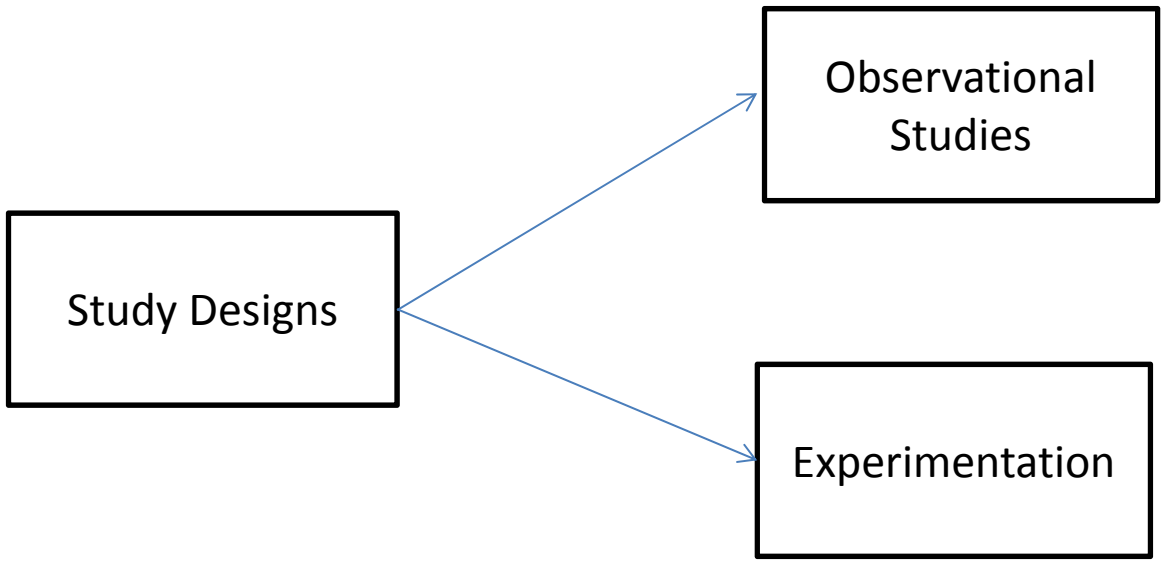
- The way a sample is selected is called the sampling plan or experimental design
- It determines the quantity of information in the sample
- It allows you to measure the reliability or goodness of your inference

REMARK

- Sampling is an important procedure. Among many available sampling plans, a sampling plan is chosen depending on the studies and available resources (available budget, time, staff).

TYPES OF SAMPLES

- Sampling can occur in two types of practical situations: Observational studies and Experimentation



TYPES OF SAMPLES

- **Observational studies**: The data existed before you decided to study or describe their characteristics. Most sample surveys in which information is gathered with a questionnaire fall into this category.
- Examples:
  - Survey of sleeping habits of students
  - For each subject record data: amount of exercise and number of colds last year
  - Ask 30 students from the class to record number of hours of daily uninterrupted sleep for a month

TYPES OF SAMPLES

- Watch for these following problems in observational studies:
  - Non-response: Are the responses biased because only opinionated people responded.
  - Under-coverage: Are certain segments of the population systematically excluded?
  - Wording bias: Has the question been poorly worded or it is too complicated?

EXAMPLE

- Example: Collecting an opinion about an issue through a radio show.
  - People who are not extremely favour or extremely against to the issue would not make the call to voice their opinion (Non-response).
  - People who never listen to a radio would not be included in the survey (Under-coverage).
  - If the issue is not explained clearly by the radio show host, then this would be a problem (Wording bias).

TYPES OF SAMPLES

- **Experimentation**: The data are generated by imposing an experimental condition or treatment on the experimental units.
- Example: Select 30 students from the class and advise 10 students to drink coffee, other 10 to tea and the other 10 to water just before bed time and ask them to record number of hours of daily uninterrupted sleep for a month.

TYPES OF SAMPLES

- Experimentation
  - Hypothetical populations: can make random sampling difficult if not possible.
  - Samples must sometimes be chosen so that the experimenter believes they are representative of the whole population.
  - Samples must behave like random samples

SAMPLING PLANS

- Note:
  - Selecting a sample design is very important step in a survey.
  - There are two different sampling exists: probability sampling and non-probability sampling.
  - Probability sampling allows us to make inferences about the population based on observation from a sample.

SAMPLING PLANS

- Probability Samples
  - Simple Random Sample
  - Stratified Sample
  - Cluster Sample
  - 1-in-k Sample
- Non-Probability Samples
  - Convenient Sample
  - Judgement Sample
  - Quota Sample

SIMPLE RANDOM SAMPLE (SRS)

- Most frequently used sampling plan is simple random sampling (SRS).
- A method of sampling that allows each possible sample of size n an equal probability of being selected is called simple random sampling

SRS - EXAMPLE

There are 89 students in a class. The instructor wants to choose 5 students to form a project group. Steps to select SRS of sample size 5

I. Give each student a number from 01 to 89
II. Choose 5 pairs of random digits from the random number table
III. Choose a random starting point in the table (any row or any column)

SRS - EXAMPLE

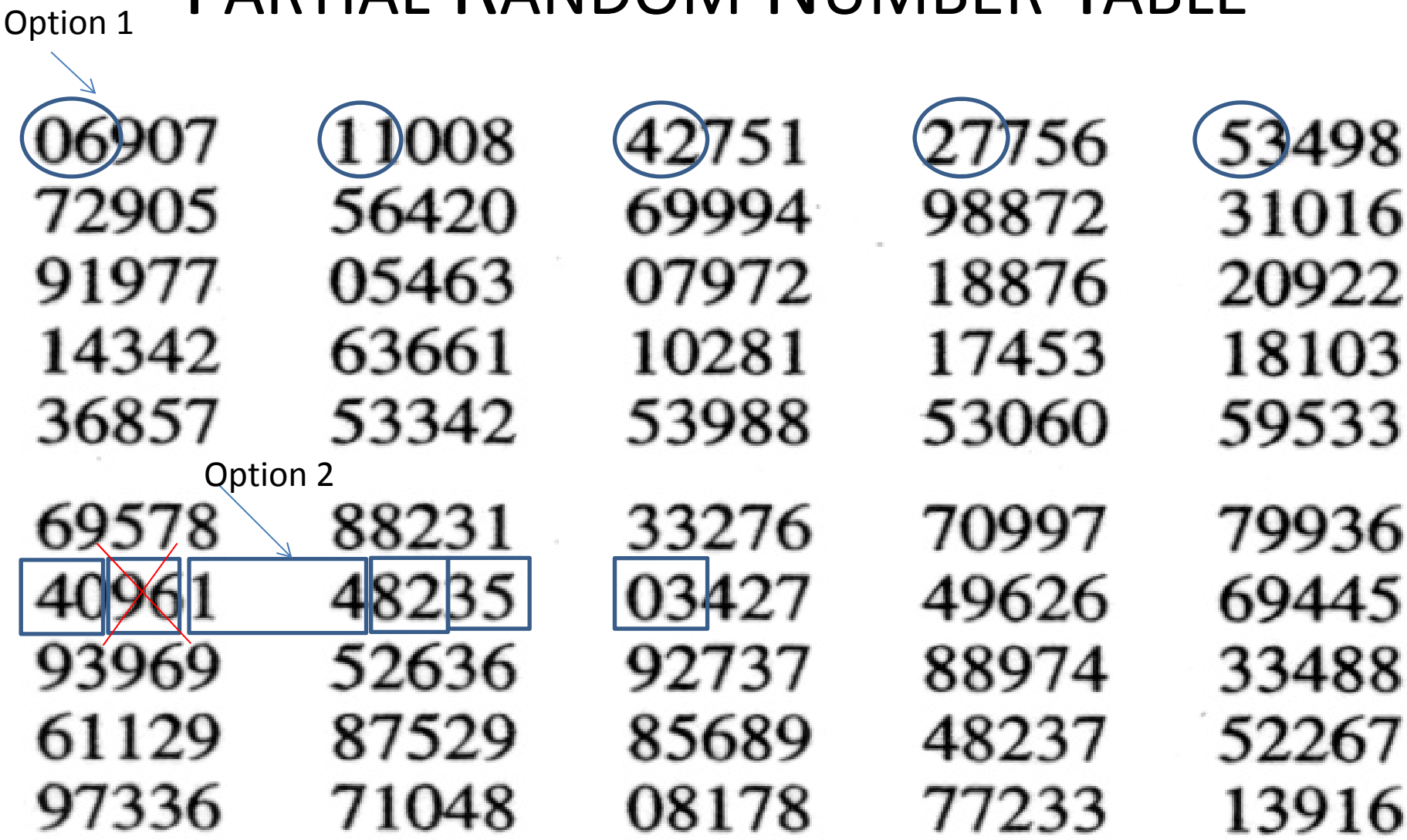
IV. If a number between 90 and 99 or 00 is chosen, choose another

V. The five students with those numbers form the group

VI. Choose the students with option 1 or option 2 selection

- Option 1: 06, 11, 42, 27, 53
- Option 2: 40, ~~96~~, 14, 82, 35, 03

Partial Random Number Table



OTHER SAMPLING PLANS

- **Stratified random sample**: divide the population into subpopulations or strata and select a simple random sample from each strata. It is more efficient than SRS
- **Cluster sample:** divide the population into subgroups called clusters; select a simple random sample of clusters and take a census of every element in the cluster. It is less precise than SRS, but more cost effective.

OTHER SAMPLING PLANS

- **1-in-k systematic sample**: Randomly select one in the r^{th}position ($1 \leq r \leq k$) of the first k elements in an ordered population, and then select every k^{th} element thereafter.
  - Units that make up the samples are in the following positions: $r, r+k, r+2k, \cdots, r+(n-1)k$

EXAMPLE

Assign numbers to each person in Canada. Select 100 random numbers using random number table or computer. 100 people with those numbers form the group	SRS
Divide Canada into provinces and take a simple random sample within each province.	Stratified
Divide Ontario into cities and take a simple random sample of 10 cities; Enumerate all people in 10 cities	cluster
Choose an entry at random from the phone book, and select every 50^{th} number thereafter	1-in-50 Systematic

NON-RANDOM SAMPLING PLANS

- There are several other sampling plans that do not involve randomization
- They should NOT be used for statistical inference!
- **Convenience sample**: a sample that can be taken easily without random selection. Examples: people walking by on the street; Students sitting in the first two of the classroom

NON-RANDOM SAMPLING PLANS

- **Judgement sample:** the sampler decides who will and won't be included in the sample
- **Quota sample:** the makeup of the sample must reflect the makeup of the population on some selected characteristic (race, ethnic origin, gender, etc.)
  - Race: Caucasian, Hispanic; Within each selected race, take a sample.

STATISTICS AND SAMPLING DISTRIBUTIONS

- Numerical descriptive measures calculated from the sample are called statistics. Example: sample mean, sample variance
- Statistics vary from sample to sample and hence are random variables.
- The probability distributions for statistics are called **sampling distributions**

STATISTICS AND SAMPLING DISTRIBUTIONS

- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs
- The probability distribution for the possible values of the statistic that results when random samples of size *n* are repeatedly drawn from the population

STATISTICS AND SAMPLING DISTRIBUTION CONT'D

- Three ways to find the sampling distribution of a statistic
  - Derive the distribution mathematically using the laws of probability (works well with small sample size)
  - Use simulation to approximate the sampling distribution empirically;
  - Use statistical theorem to derive exact or approximate sampling distributions.

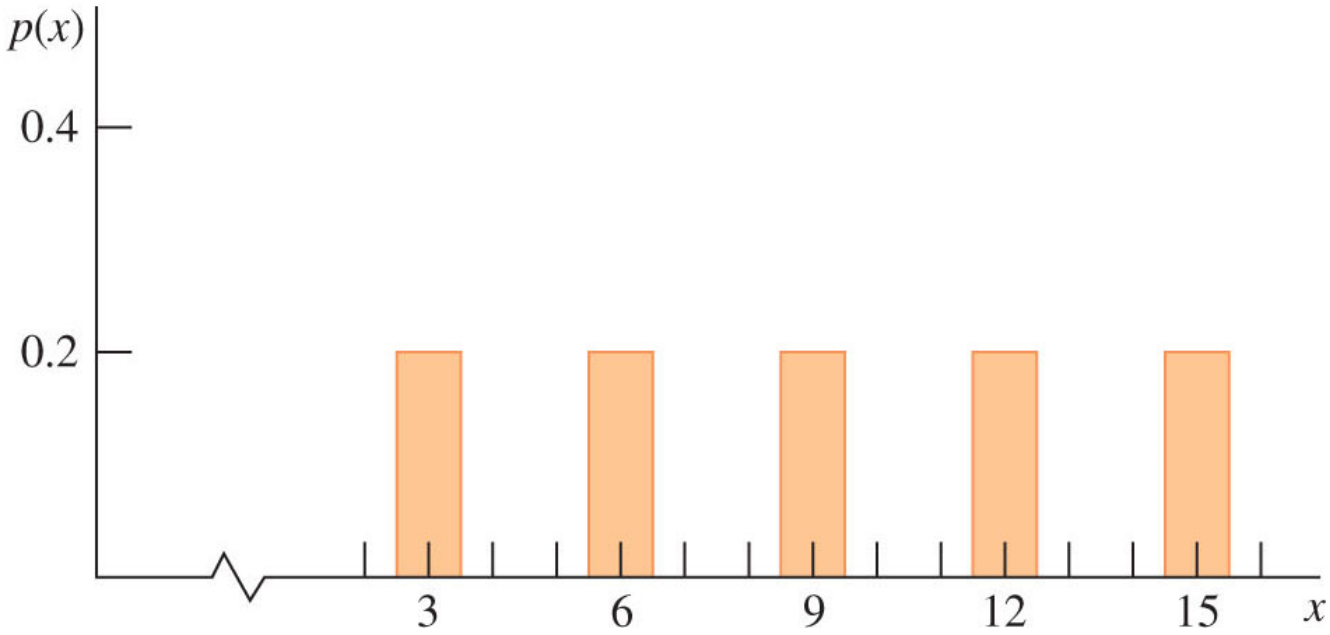
EXAMPLE

- Deriving the distribution mathematically using the law of probability
- Population: 3, 6, 9, 12, 15. Draw samples of size n=3 without replacement.

$$population\ median = 9$$

$$Population\colon\ \mu = \frac{3 + 6 + 9 + 12 + 15}{5} = 4.5$$

PROBABILITY HISTOGRAM



EXAMPLES CONT'D

- All possible samples, their mean and median

Sample	Sample Values	$\bar{x}$	m
1	3, 6, 9	6	6
2	3, 6, 12	7	6
3	3, 6, 15	8	6
4	3, 9, 12	8	9
5	3, 9, 15	9	9
6	3, 12, 15	10	12
7	6, 9, 12	9	9
8	6, 9, 15	10	9
9	6, 12, 15	11	12
10	9, 12, 15	12	12

EXAMPLE –SAMPLING DISTRIBUTIONS

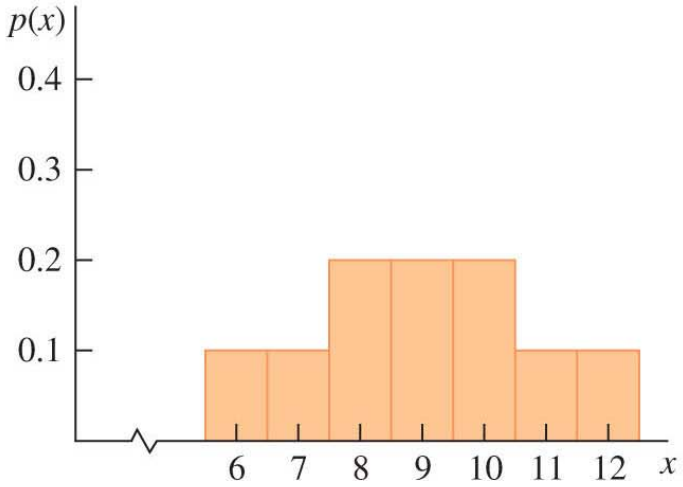
(a) $\bar{x}$	$p(\bar{x})$
6	0.1
7	0.1
8	0.2
9	0.2
10	0.2
11	0.1
12	0.1

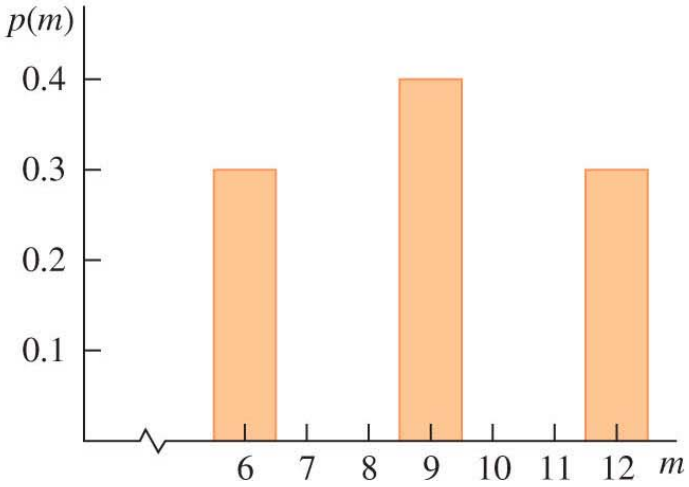
(b) m	$p(m)$
6	0.3
9	0.4
12	0.3

EXAMPLE – SAMPLING DISTRIBUTIONS

$$P(\bar{x} = 8) = \frac{2}{10} = 0.2$$

$$P(m = 6) = \frac{3}{10} = 0.3$$





REMARKS

- As the population is small, it is easy derive sampling distribution using laws of probability
- Take population size = N = 25, sample size = n = 5. Number of possible samples = $C_5^{25} = 5313$. Too many samples to derive sampling distribution using laws of probability.