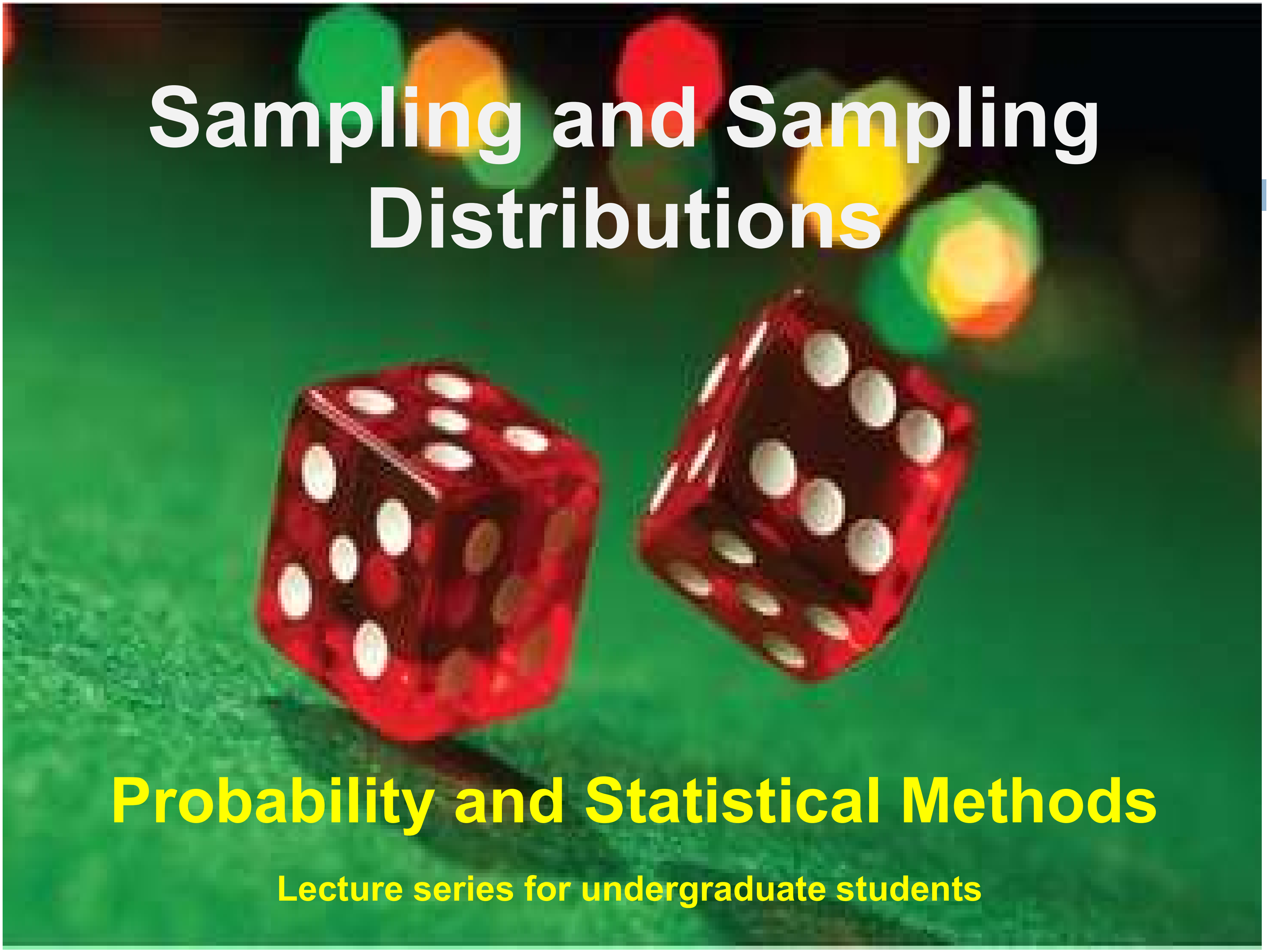


Sampling and Sampling Distributions

Probability and Statistical Methods

Lecture series for undergraduate students

Sampling and Sampling Distributions

The objective of most statistical investigations is inference- that is, making decisions or predictions about a population based on information obtained in through sample.

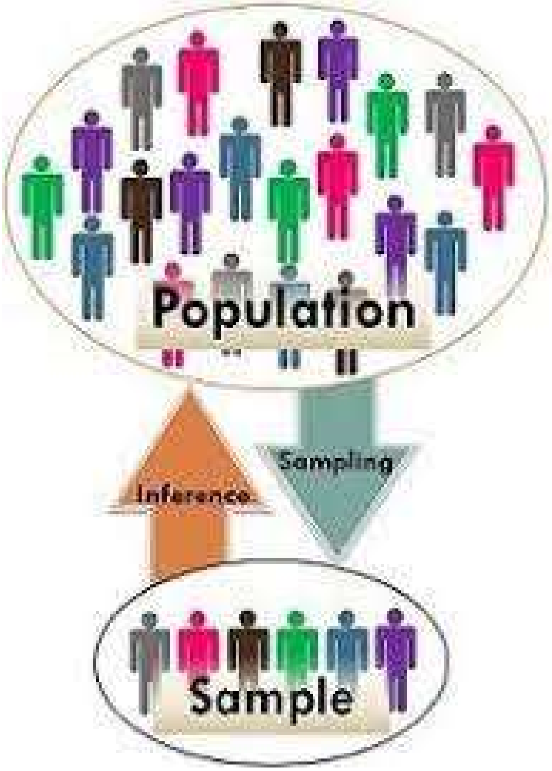
Lecture 05

Week-05

Dr. Tahseen A. Jilani

Basic Terminology

- A **population** is the collection of all the elements of interest.
- A **sample** is a subset of the population.
- **Sampled population** – Population from which sample drawn. Researcher should clearly define.
- **Frame** – list of elements that sample selected from. e.g. telephone book, city business directory.



Basic Terminology

- □ **Parameters** are characteristics of a **target population.** Parameters may also be termed **population values.**

- □ A **statistic** is also referred to as a **sample statistic** or, when estimating a parameter, a **point estimator** of a parameter. A specific value of a point estimator is referred to as a **point estimate** of a parameter. e.g sample mean $(\bar{x}\ or\ \hat{\mu})$, sample variance $(s^2\ or\ \hat{\sigma}^2)$, sample proportion (p or $\hat{P}$), sample ratio (r or $\hat{R}$) etc. to estimate (or to make conclusions) about ***population parameter*** such as population mean $(\mu$, population variance σ^2, population proportion (P), population ratio (R) etc.

Statistical Inference/Statistical Decision Making

- **Sampling distribution** of a statistic is the probability distribution of that statistic.
  - For example, sampling distribution of sample means is normal distribution.
  - Sampling distribution of sample proportions is again normal distribution
  - Sampling distribution of single variance is sample Chi-Square distribution.

  And so on.

Why sample information is so useful in decision making

- There could be many samples if a population is large.
  - e.g. 1% of population N = 1,000,000 is 10,000.
  - e.g. COVID19 cases in whole Pakistan is 5% based on a sample data of 40,000 covid19 tests. Sample results used to estimate P.
  - costly due to time and other resources required).
- For large populations, a carefully collected data could produce more than 95% same results (as if a complete census is performed.)
- If there could be ${}^{N}C_{n}$ possible different samples, then the sample characteristic (the statistic e.g. sample mean $\bar{x}$) is a random variable and would have its own probability distribution.

Sample from a process

- It may be difficult or impossible to obtain or construct a frame.
  - Larger or potentially infinite population – fish, trees, manufacturing processes.
  - Continuous processes – production of milk or other liquids, transporting commodities to a warehouse.
- Random sample is one where any element selected in the sample:
  - Is selected independently of any other element.
  - Follows the same probability distribution as the elements in the population.
- Careful design for sample is especially important.
  - Sample production of milk at random times.
  - Forest products – randomly select clusters from maps or previous surveys of tree types, size, etc.

Sampling Techniques

- □ **<u>Random sampling methods (SRS)</u>** – each member has an equal probability of being selected.
- □ **<u>Stratified samples</u>** – sample from each stratum or subgroup of a population. E.g. region, size of firm.
- □ **Systematic** – every *k*th case. Equivalent to random if patterns in list are unrelated to issues of interest. e.g. telephone book.
- □ **Cluster samples** – sample only certain clusters of members of a population. E.g. city blocks, firms.
- □ **Multistage samples** – combinations of random, systematic, stratified, and cluster sampling.

Sample selection from a hypothetical Population

- *N* is the symbol given for the size of the population or the number of elements in the population.
- *n* is the symbol given for the size of the sample or the number of elements in the sample.
- **Simple random sample (SRS)** is a sample of size *n* selected in a manner that each possible sample of size *n* has the same probability of being selected.
- In the case of a random sample of size *n*=*1*, each element has the same chance of being selected.

Sample selection methods

- **Sampling with replacement** – after any element randomly selected, replace it and randomly select another element. But this could lead to the same element being selected more than once.

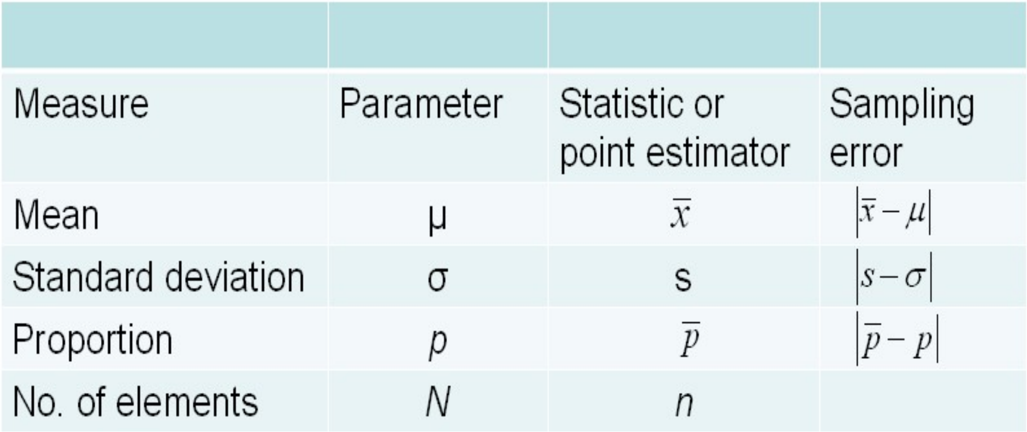
$$m = N^n$$

- More common to **sampling without replacement.** Make sure that on each stage, each element remaining in the population has the same probability of being selected.

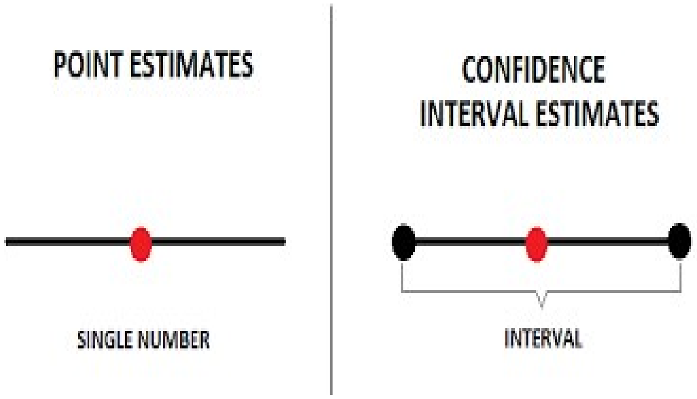
$$m = C_n^N = \frac{N!}{n!(N-n)!}$$

Point Estimation

- The proportion is the frequency of occurrence of a characteristic divided by the total number of elements.
- The proportion of elements of a population that take on the characteristic is *p* and the proportion of the elements in the sample selected with this same characteristic is .



Measure	Parameter	Statistic or point estimator	Sampling error
Mean	μ	$\bar{x}$	$\lvert\bar{x}-\mu\rvert$
Standard deviation	σ	s	$\lvert s-\sigma\rvert$
Proportion	p	$\bar{p}$	$\lvert\bar{p}-p\rvert$
No. of elements	N	n	



Example: Sampling without replacement

Consider a hypothetical population consisting of five numbers **0, 2, 4, 6 and 8** (N=5). Select all possible random samples of size n = 2 from the given population. (A). Without replacement and (B). With replacement

For without replacement case: $^{5}C_{2}=10$

For with replacement case: $(5)^{2}=25.$

The population mean is $\mu = \frac{0+2+4+6+8}{5} = 4.00$

Example: Sampling without replacement

let the population is 0, 2, 4, 6, 8 (only N=5) here let n =2

Sampling without Replacement

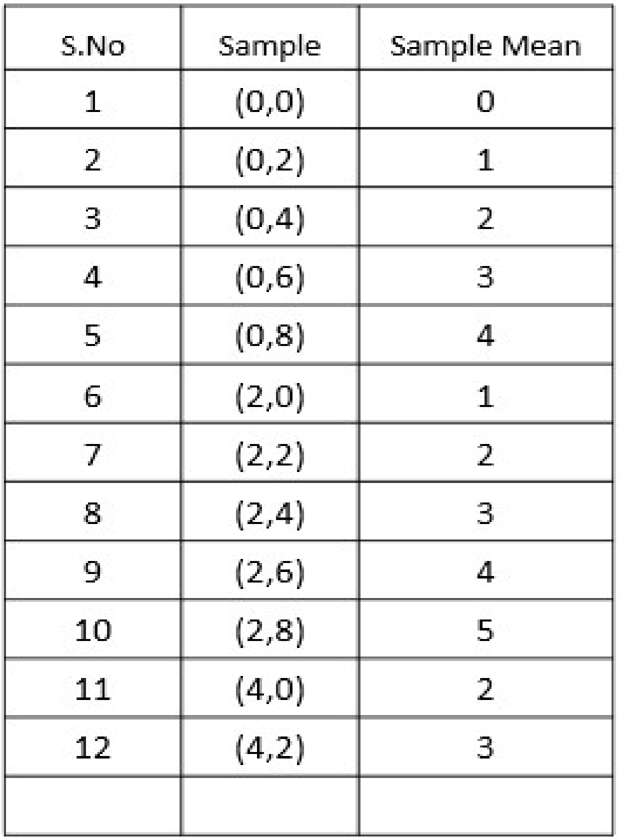
S.No	Sample	Sample Mean
1	(0,2)	(0+2)/2 = 1
2	(0,4)	(0+4)/2 = 2
3	(0,6)	(0+6)/2 = 3
4	(0,8)	(0+8)/2 = 4
5	(2,4)	3
6	(2,6)	4
7	(2,8)	5
8	(4,6)	5
9	(4,8)	6
10	(6,8)	7

He sampling distribution for Sample mean is:

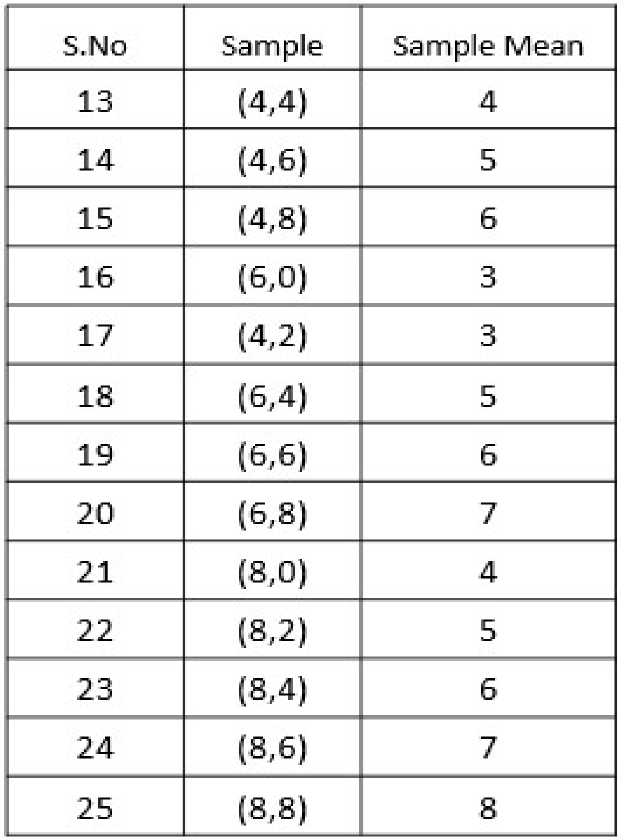
$\bar{x}$	1	2	3	4	5	6	7
P($\bar{x}$)	.1	.1	.2	.2	.2	.1	.1

As the sample mean values are changing so we can Say it's a random variable and so we can define its probability distribution.

Example: Sampling with replacement $m = N^n$



S.No	Sample	Sample Mean
1	(0,0)	0
2	(0,2)	1
3	(0,4)	2
4	(0,6)	3
5	(0,8)	4
6	(2,0)	1
7	(2,2)	2
8	(2,4)	3
9	(2,6)	4
10	(2,8)	5
11	(4,0)	2
12	(4,2)	3



S.No	Sample	Sample Mean
13	(4,4)	4
14	(4,6)	5
15	(4,8)	6
16	(6,0)	3
17	(4,2)	3
18	(6,4)	5
19	(6,6)	6
20	(6,8)	7
21	(8,0)	4
22	(8,2)	5
23	(8,4)	6
24	(8,6)	7
25	(8,8)	8

DIY - Prepare sampling distribution for sample means in this example

Sampling Distribution

- □ A **sampling distribution** is the probability distribution for all possible values of the sample statistic (e.g. sampling distribution for sample mean).
- □ Each sample contains different elements so the value of the sample statistic differs for each sample selected. These statistics provide different estimates of the parameter. The sampling distribution describes how these different values are distributed.
- □ For example: With the sampling distribution of $\bar{x}$, we can “make probability statements about how close the sample mean $\bar{x}$ is to the population mean μ”.

Sampling distribution of the sample mean

- When a sample is selected, the sampling method may allow the researcher to determine the sampling distribution of the sample mean. The researcher hopes/expects that the mean of the sampling distribution will be μ (which is unknown). Thus $E(\bar{x}) = \mu$, that is that sample mean is an unbiased estimator of μ.
- If the variance of the sampling distribution of $\bar{x}$ can be determined, then the researcher is able to determine how random variable $\bar{x}$ is close to μ.
- The researcher hopes to have a small variability for the sample means, so most estimates of μ are close to μ.

Sampling distribution of the sample mean

- □ If a simple random sample is drawn from a normally distributed population, the sampling distribution of mean is normally distributed.
- □ The mean of the sampling distribution of $\bar{x}$ sample mean is μ, the population mean $\mathrm{E}(\bar{x}) = \mu$.
- □ If the sample size *n* is reasonably large proportion of the population, then the variance of sample mean is$\mathrm{V}ar(\bar{x}) = \frac{\sigma^2}{n}$.
- □ Here the population SD is given which is not known mostly.

Point Estimation

- Note: If n/N > 0.05, it may be best to use the finite population correction factor.

- Sampling With replacement $E(\bar{x}) = \mu \; and \; V(\bar{x}) = \frac{\sigma^2}{n}$

- Sampling Without replacement $E(\bar{x}) = \mu \; and \; V(\bar{x}) = \frac{\sigma^2}{n}(\frac{N-n}{N-1})$

	Normally distributed population	Sampling distribution of □x when sample is random
No. of elements	N	n
Mean	μ	μ
Standard deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

CENTRAL LIMIT THEOREM

[illegible] with your [illegible] the previous slides, or to refresh a bit, or to ask questions.

Administrative Issues

Groups Formation

Central Limit theorem (C.L.T.)

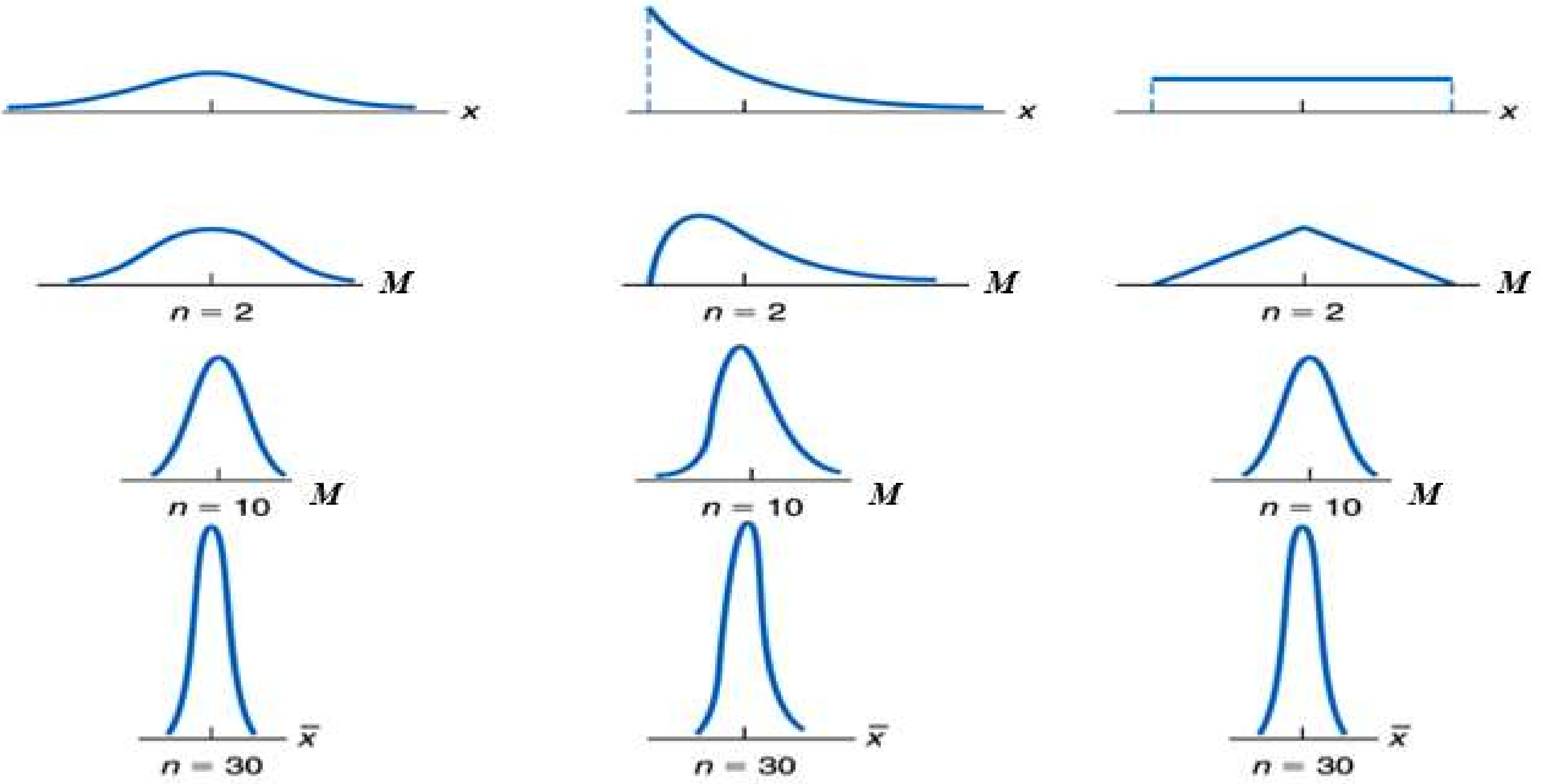
- When sampling is done from a population with mean μ and finite SD = σ, the sampling distribution of $\bar{x}$ will tend to a normal distribution with mean $E(\bar{x}) =$ μ and SD($\bar{x}$) = $\frac{\sigma}{\sqrt{n}}$ (also termed as **standard error** of the sample mean) as the sample size *n* becomes large.

- If the population from which the sample is drawn is symmetrically distributed, *n* > 30 may be sufficient to use the CLT.

The sampling distribution for $\bar{x}$ ~ N(μ, ${}^{\sigma^2}/_{n}$)

Central Limit theorem (C.L.T.)

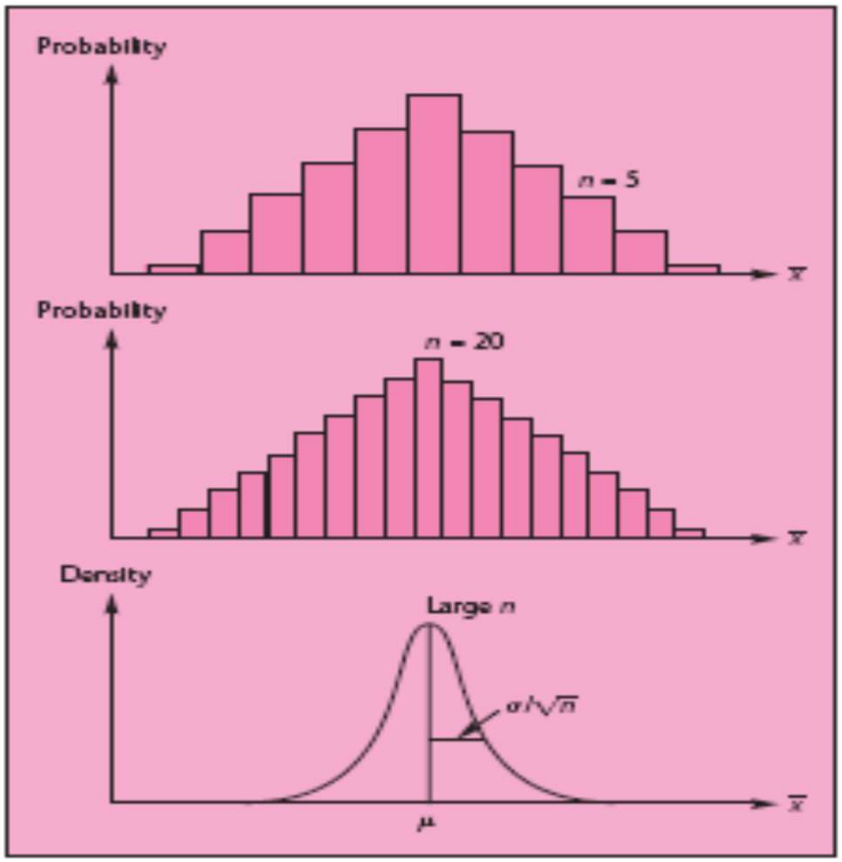
- □ In general, a sample of 30 or more elements is considered **large enough** for the central limit theorem to take effect.



Central Limit theorem (C.L.T.)

- The CLT is remarkable because it states that the distribution of the sample mean tends to a normal distribution *regardless* of the distribution of the population from which the random sample is drawn.
- CLT pays central role in statistical decision making.

The Sampling Distribution of $\overline{X}$ as the Sample Size Increases



2. The Standardized Sampling Distribution of the Sample Mean $\bar{x}$ When σ Is Unknown but n is large (n >30)

- The sampling distribution for $\bar{x}$ ~ N(μ, ${}^{\sigma^2}/_{n}$)
- To use the central limit theorem, we need to know the population standard deviation, σ. When σ is not known, we use its estimator, the sample standard deviation *S*, in its place. In such cases, the distribution of the standardized statistic

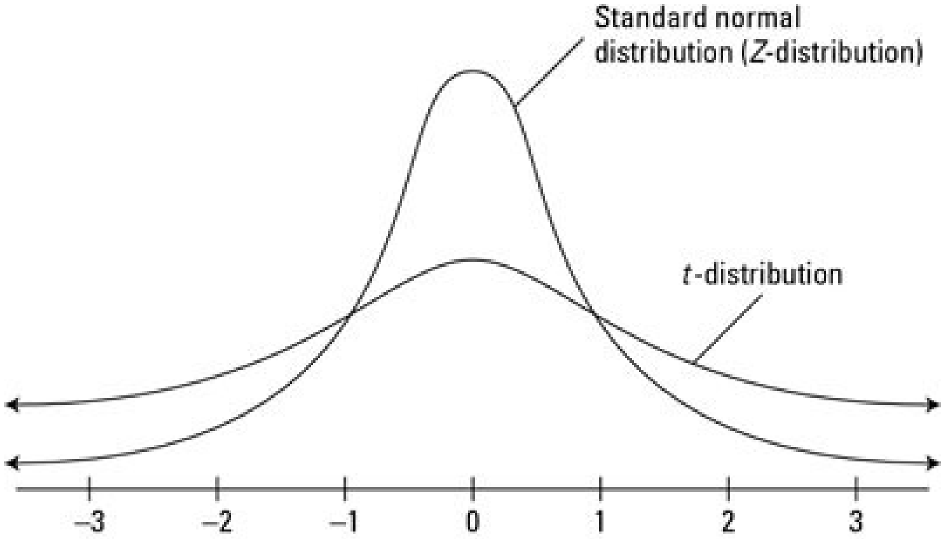
$$z = \frac{\bar{x} - \mu}{{}^{s}/_{\sqrt{n}}}$$

(where *s is an unbiased estimator of* σ) **<u>provided that n is large</u>**.

3. The Standardized Sampling Distribution of the Sample Mean $\bar{x}$ When σ is unknown but n is small (n<30)

□ The distribution of sample mean is no longer a normal distribution. And is replaced by students t-distribution with *(n-1) degrees of freedom*. The students' *t*-distribution has wider tails than the standard normal distribution.

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$



Degrees of freedom

- **The number of free observations in a dataset.**
  - Suppose you are asked to choose 10 numbers. You then have the freedom to choose 10 numbers as you please, and we say you have 10 **degrees of freedom.**
  - But suppose a condition is imposed on the numbers. The condition is that the sum of all the numbers you choose must be 100. In this case, you cannot choose all 10 numbers as you please. After you have chosen the ninth number, let's say the sum of the nine numbers is 94. Your tenth number then has to be 6, and you have no choice. Thus you have only 9 (i.e. 10-1) degrees of freedom.
- In general, if you have to choose *n* numbers, and a condition on their total is imposed, you will have only (*n-1*) degrees of freedom.

Example - Central Limit theorem (C.L.T.)

- Given a population distribution with $\mu = 32$ and $\sigma = 12$. What is the probability of drawing a sample of size n = 36 where $\bar{x} > 34$.
- Here σ *is known (so use z distribution).*

$$P(\bar{x} > 34) = 1 - P(\bar{x} < 34) = 1 - P\left(z < \frac{34 - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= 1 - P\left(z < \frac{34 - 32}{\frac{12}{\sqrt{36}}}\right) = 1 - P\left(z < \frac{2 \times 6}{12}\right)$$

$$= 1 - P(z < 1.00) = 1 - 0.8413 = 0.1587$$

Examples: Central Limit theorem

- □ In a distribution with $\mu = 45$ and $\sigma = 15$ what is the probability of drawing a sample of n=25 with $\bar{x}$ >50?

Here σ *is know so use Z distribution.*

$$P(\bar{x} > 50) = 1 - P(\bar{x} < 50) =$$

- □ In a distribution with $\mu = 90$ and $\sigma = 18$, for a sample of $n =$ 36, what sample mean $\bar{x}$ would constitute the boundary of the most extreme 5% of scores?

Further examples on Central limit theorem (CLT)

- A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean μ = 205 pounds and standard deviation σ = 15 pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

- https://www.webassign.net/question_assets/idcollabstat2/Chapter7.pdf

- https://gtribello.github.io/mathNET/central-limit-theorem-problems.html

- http://www.stat.ucla.edu/~nchristo/introeconometrics/introecon_central_limit_theorem.pdf

Questions (to do): Central Limit theorem

- Analysis of commuter travel shows that the number of passenger per car, X, is a discrete random variable with independent, identical distributions, such that E(X)=1.2 and var(X)=1.0. Use the central-limit theorem to estimate the probability that, in a sample of n=100 cars, the total number of passengers is 140 or fewer.

- The share price of SOR plc varies in a random manner, such that the price increase each minute is described by a discrete random variable X (measured in pounds), with the following probability mass

$$f_X(x) = \begin{cases} 0.5 & , \quad x = +0.05 \\ 0.2 & , \quad x = 0.00 \\ 0.3 & , \quad x = -0.05 \end{cases}$$

- Calculate E(X)=μE(X)=μ and √var(X)=σvar(X)=σ and hence use the central limit theorem to estimate the probability that the price will increase by £1.20 , or more, after 3 hours.

Relationship between sample size and SE($\bar{x}$) in Central Limit theorem

- □ In SD($\bar{x}$) when sample size increase then the value of SD($\bar{x}$) decrease.

$$V(\bar{x}) = \sigma^2/_n$$

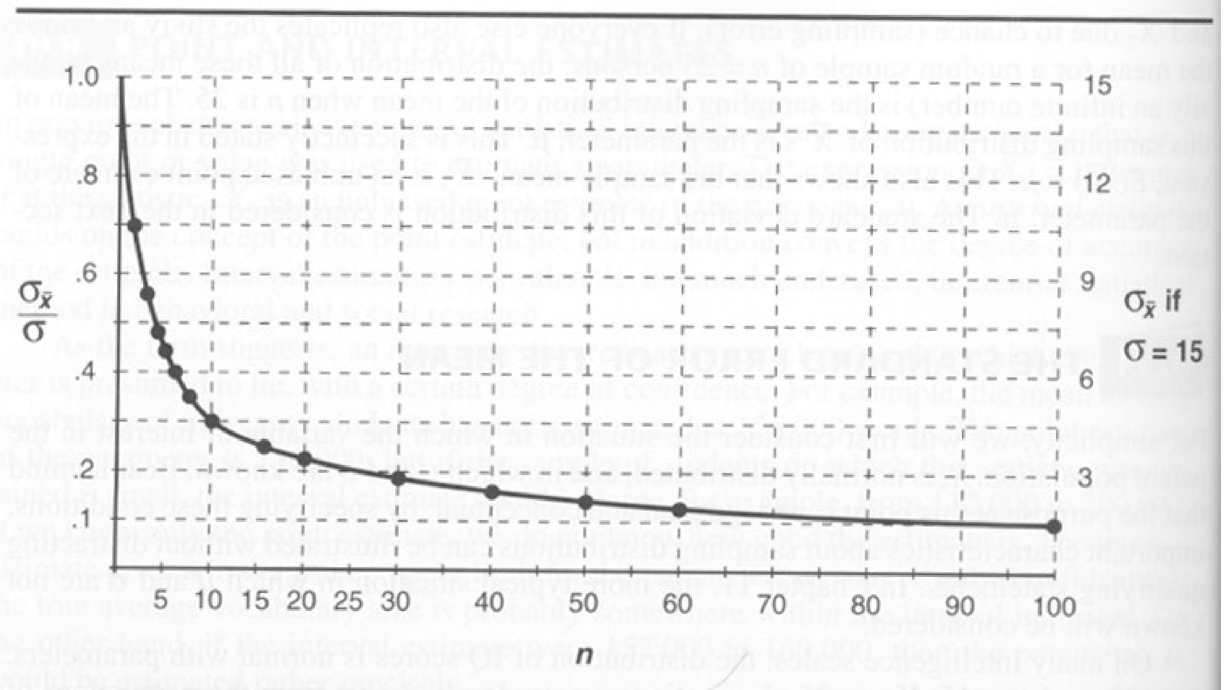


FIGURE 10.1 The Relationship of *n* and $\sigma_{\bar{x}}$

End of Lecture

Week-05