Sampling and

Sampling Distribution

S CAgarkar

VN BRIMS, Thane

Some Experiences

The rating of a Television programme is decided by getting feedback from a few

selected houses.

Exit pollopinion is given by asking a few voters in

The locality.

A qualityof sweet is decided by tasting a small

amount of it.

The trendof a product sell is determined by getting

Data from a few sales outlets.

.

Census vs Sampling

Every ten years the census is conducted in our Country. It involves the collection of information From allthe individuals. People appointed for this Purpose gofrom door to door and get detailed Information aboutevery member of the family.

On the other hand studies are conducted by drawing a small number of cases by ensuring that itrepresents the entire population faithfully.

.

National Sample Survey

Government of India has set upanorganization

Named NationalSample Survey (NSS) which

Conducts surveyson a variety of issues in the

Country everyyear.

Based on the data made available by the NSS the

Planning Commission (nowNITYAyog) plans a

Variety ofdevelopmental activities for the country.

The information about the work of NSS canbe found at thewebsite of the organization.

Some Terms

- **Population**: Statisticians use the term population to refer to people or items chosen for study.
- **Sample**: Statisticians use the term sample to describe a portion chosen from the population.
- **Statistics and Parameters**: A parameter is a characteristics of population while a statistic is a characteristics of sample.
- We can use statistic to estimate the parameter. For example height of tenth graders from a school to guess the height of tenth graders in the entire country.

Some symbols

	Population	Sample
Definition	Collection of items being considered	Portion of the population chosen for study
Characteristics	Parameters	Statistic
Symbols	Population size=N	Sample size = n
	Population mean =μ	Sample mean = x
	Standard deviation =σ	Standard deviation = s

The Need for Sampling

- It is difficult to get the entire population for study. Hence one tries to get a reasonable sample. For example all Indians to comment on Delhi verdict is impossible to get. Instead we try to get a sample from Thane.
- It is difficult to handle the population as the data generated is huge.
- Dealing with entire population is costly and time consuming.
- It is not necessary to deal with the entire population. Studies made on a small sample can be generalized with reliability.

Types of Sampling

- The essential characteristics of the sample is that it should be representative of the population from which it is drawn. To ensure this two types of methods are followed in selecting a sample.
- **NonRandom orjudgement sampling**: Personal knowledge and opinion are used to identify those items from population that are to be included in the sample. For example, trees in the forest.
- **Random or Probability Sampling**: In this method all items in the population have a chance of being chosen in the sample.

Judgment Sampling

- In judgment sampling personal knowledge and opinion are used to identify the items from the population.

- A sample selected by judgment sampling is based on someone's experience about the population.

- Sometimes a judgement sampling is used as a pilot or trial sample to understand the issue at hand or to decide how to take random sample later.

Random Sampling

- **Simple Random Sampling**: It is done either by the use of random number table or by random choosing of slips.
- **Systematic Sampling**: Choose the person or item with fixed number and then following the fixed interval. For example every tenth candidate.
- **Stratified Sampling**: Divide population into homogeneous strata then choose sample using above methods. For example, people from different age groups, or products of different days.

Random Sampling Cont.

- **Cluster Sampling**: Divide population into groups or clusters. For example to know the amount of moneyspentbyganeshmandalsfromthecity divide the city in different wards and then take the sample from a certain ward.
- Stratified sampling and cluster sampling may look alike. Stratified sampling is adopted when each group has small variation within itself but there is a wide variation between the groups. Cluster sampling on the other hand is adopted when there is a considerable variation within each group but groups are essentially similar.

Examples

- If we have a population of10,000 andwe wish to sample 20 randomly, use the random digittableto select 20 individuals from 10000.Listthe numbers of elements selected based on the random number digit table.

- We wish to sample 15 pages from this textbook. Use the random number table to select 15 pages and count the number of words in italics on each page. Report your result.

Design of Experiments

- Many times we need to justify the claims made of products. Let us for example, assume that a new battery producing company claims higher life for its battery.

- In this case we need to conduct an experiment by selecting a representative sample of new and old batteries and test them under conditions that are similar for both categories.

- Keeping other conditions same and varying just one is called controlled experiments.

Reacting to Claims

- Many times the advertisers make high claims. They are to be taken with a pinch of salt.

- Look for the advertisements on television, radio, and newspapers and find out those that make high claims.

- Design an experiment to testify the claims keeping other parameters constant.

Sampling Distribution

- A probability distribution of all the possible means of the samples is a distribution of sample means. Statisticians call this as a**Sampling Distribution of the mean**.
- We can also have a**sampling distribution of a proportion**. Assume that we have determined the proportion of beetle infected pine trees in a sample of 100 trees taken from a very large forest. We have taken a large number of those 100-item samples. A plot of probability distribution of the possible proportions of infested sample will give distribution of sample proportions.

Sampling Distribution

Population	Sample	Sample Statistic	Sampling Distribution
Water in a river	10 gallon container of water	Mean of Mercuryin ppm ofwater	Sampling distribution of the mean
Allprofessional basketballplayer	Groups of five players	Median height	Sampling distribution of the median
All parts produced by a manufacturing process	50 parts	Proportion defective	Sampling distribution of the proportion

The Standard Error

- The standard deviation of the distribution of a sample statistic is known as the standard error of the statistic.
- The term standard error is used because it conveys a specific meaning. Forexample,the height of a freshman in a large university. We could take a series of samples and calculate the mean height. We expect a some variability in our observed means. This variability results from sampling error due to chance.

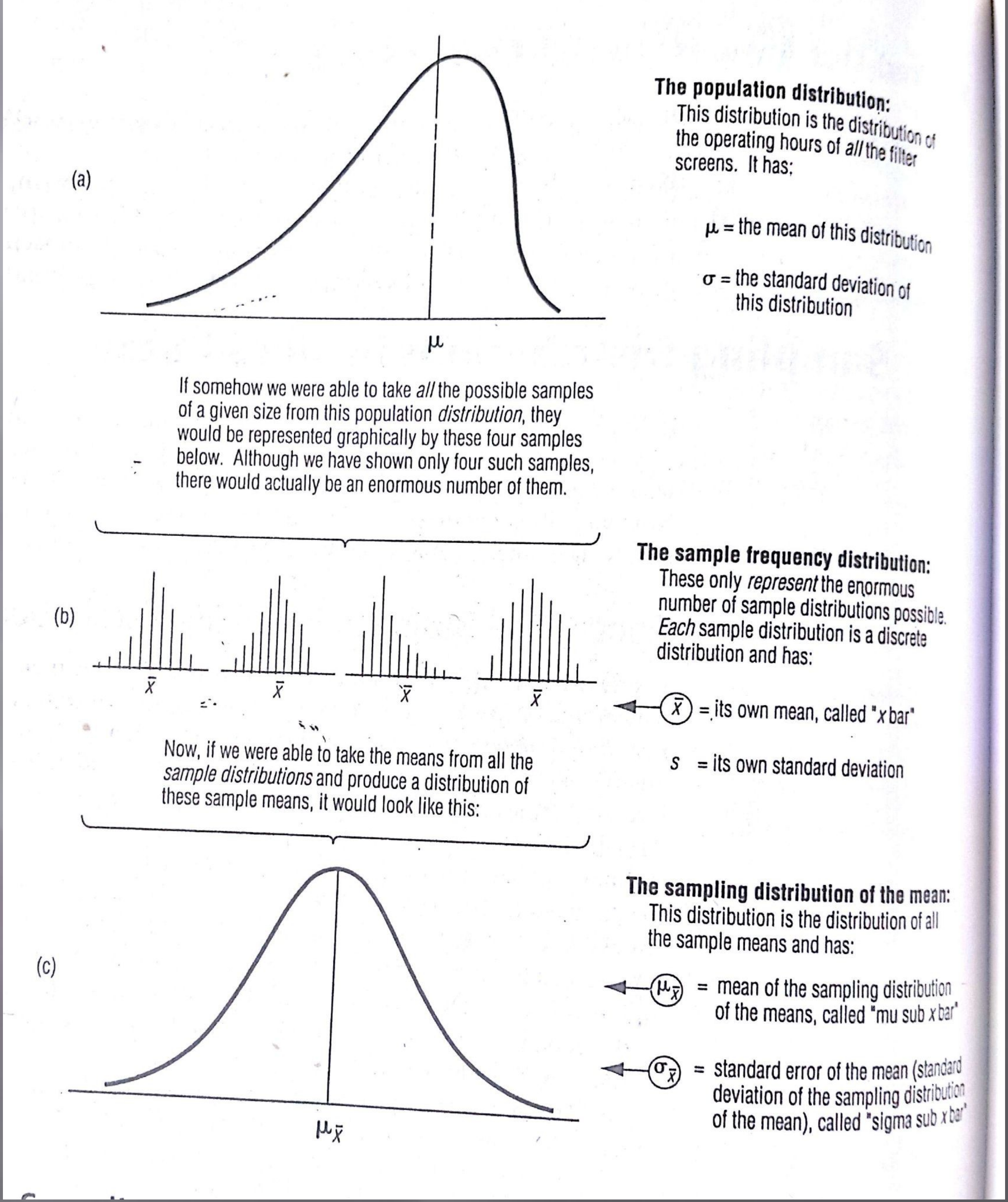
The terminology

When we wish to refer to the	Weuse theconventional term
Standard deviation of the distribution of sample means	Standard error of the mean
Standard deviation of the distribution of sample proportions	Standard error of the proportion
Standard deviation of the distribution of sample medians	Standard error of the median
Standard deviation of the distribution of sample ranges	Standard error of the range

Conceptual Basis

- We willget three distributions: The population distribution, The sample frequency distribution and The sampling distribution of mean. They are different fromeach other.

- Assume that we are studying the distribution of operating hours of all the filter screens in a large industrial pollution control system before a screenbecomecloged. Threedistributions associated are shown in the following figure.



The population distribution:
This distribution is the distribution of the operating hours of *all* the filter screens. It has:

μ = the mean of this distribution

σ = the standard deviation of this distribution

If somehow we were able to take *all* the possible samples of a given size from this population *distribution*, they would be represented graphically by these four samples below. Although we have shown only four such samples, there would actually be an enormous number of them.

The sample frequency distribution:
These only *represent* the enormous number of sample distributions possible. *Each* sample distribution is a discrete distribution and has:

$\bar{X}$ = its own mean, called "*x* bar"

s = its own standard deviation

Now, if we were able to take the means from all the *sample distributions* and produce a distribution of these sample means, it would look like this:

The sampling distribution of the mean:
This distribution is the distribution of all the sample means and has:

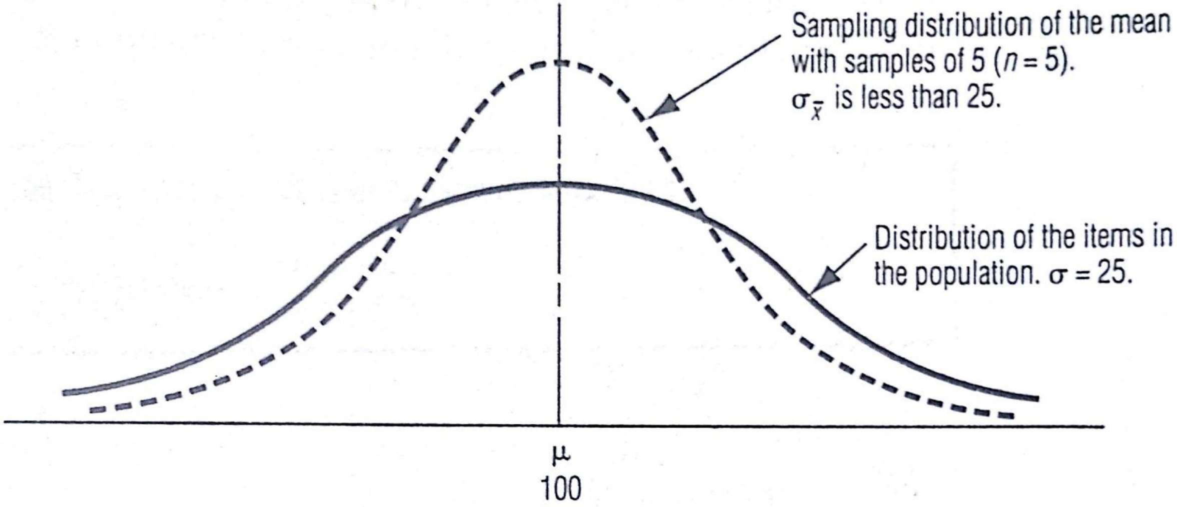
$\mu_{\bar{x}}$ = mean of the sampling distribution of the means, called "mu sub *x* bar"

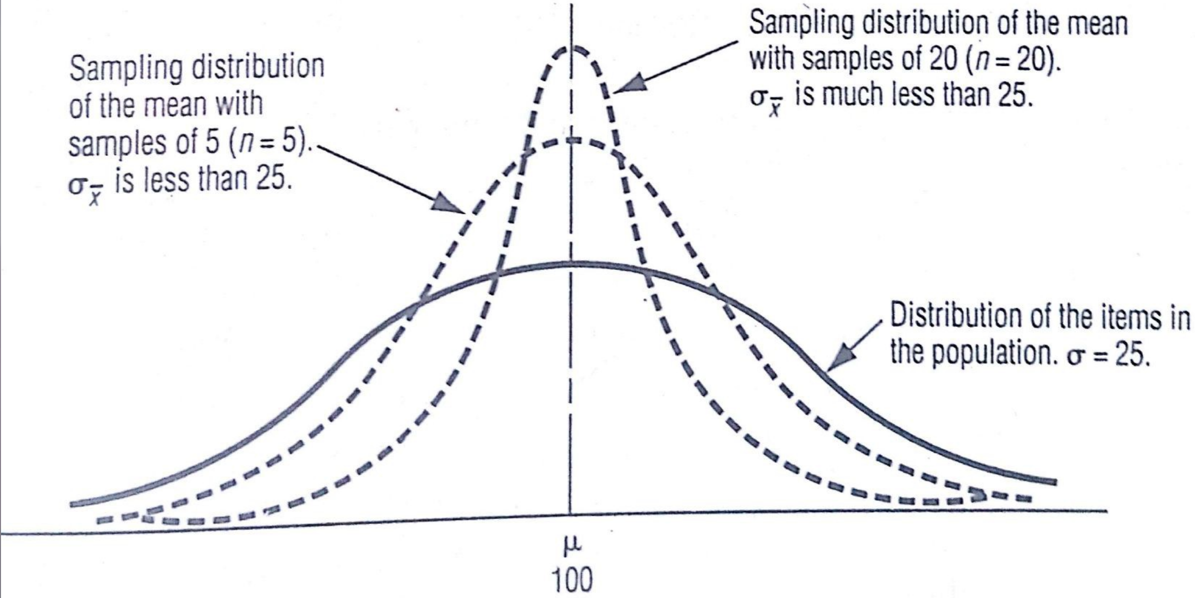
$\sigma_{\bar{x}}$ = standard error of the mean (standard deviation of the sampling distribution of the mean), called "sigma sub *x* bar"

Sampling fromnormal distribution

- Suppose we draw samples from a normally distributed population. In thiscasewe find thatthe samplingdistribution has a mean equal to the population mean ($\mu_X = \mu$).

- The sampling distribution has a standard deviation (a standard error) equal to the population standard deviation divided by the square root of the sample size.
- $SE = \sigma \div \sqrt{n}$

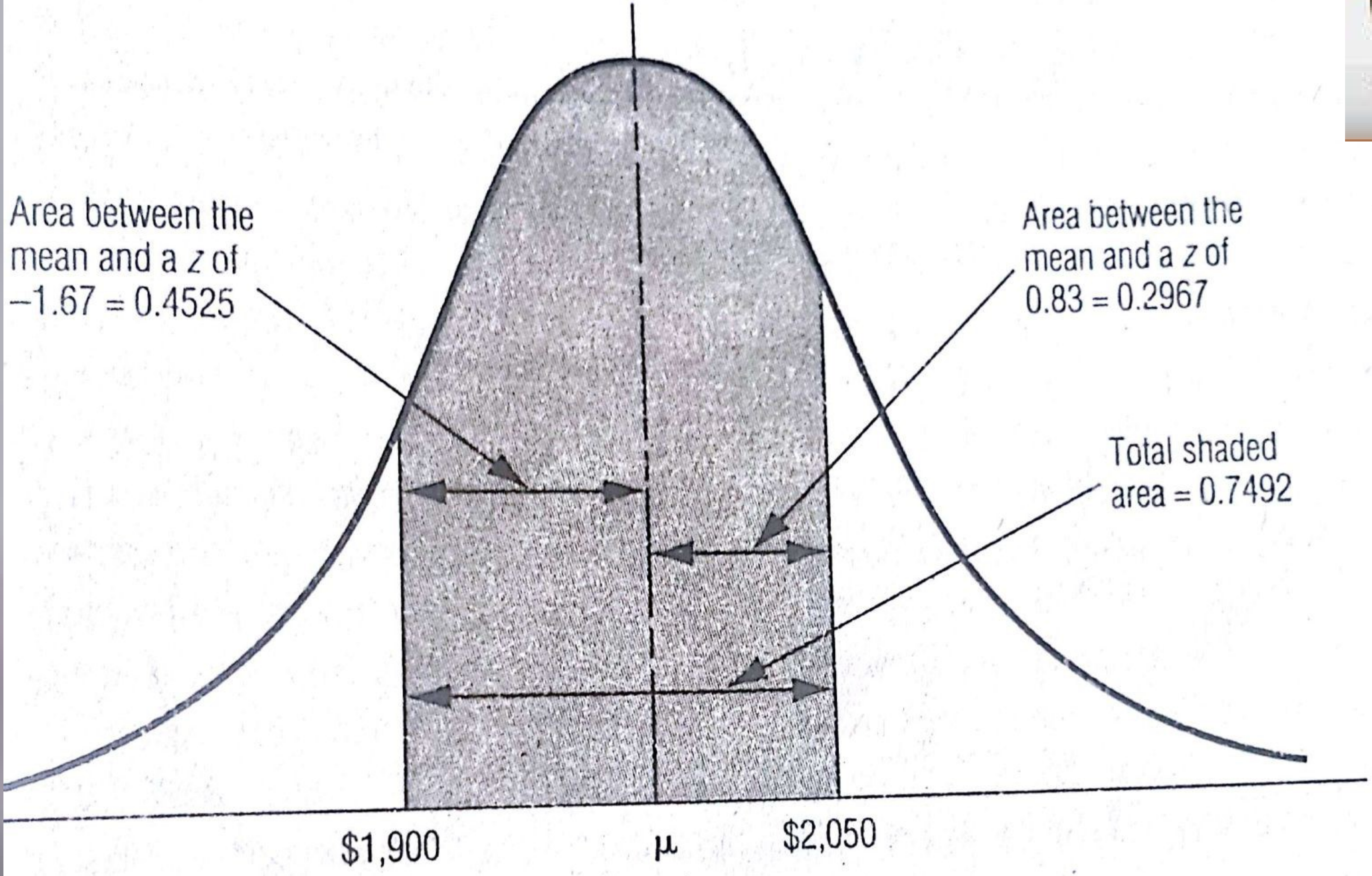




size from 5 to 20. This would not ch

Example 1

- A bank calculates that its individual savings accounts are normally distributed with a mean of $ 2,000 and a standard deviation of $ 600. If a bank takes a random sample of 100 accounts, what is the probability the sample mean will lie between $ 1,900 and $ 2,050?
- Standard Error of the mean $SE = \sigma \div \sqrt{n} = 60$
- Z value = Difference $(x - \mu)/SE$
- For $x=1900$, $z = -1.67$
- For $x= 2050$, $z = 0.83$
- Z value of -1.67 means an area of 0.4525 while the z value of 0.83 means an area of 0.2967.
- If we add these values we get 0.7492 as the total probability that the sample mean will be between $1900 and 2050. This result can also be shown graphically (p 317).



6.5 **Sampling Distributions in More Detail**

Example 2

- Daily sales figures of 40 shopkeepers were calculated and the average sales and S.D. were found to be Rs. 528 and 600 respectively. Is the assertion that daily sales on the average is Rs. 400 only contradicted at 5% level of significance by the sample?
- SE = 94.94
- Z = Difference/SE = 1.348
- The value is less than 1.96 (as per the table at 5% level of significance). Therefore, the hypothesis is accepted.

Problem to solve

- An Auditor of a large credit card company knows that on an average the monthly balance of any given customer is $112 and the standard deviation is $56. if the auditor audits 50 randomly selected accounts, what is the probability that the sample average monthly balance is

- A) below $100
- B) between $100 and $130.
- Ans: 0.0643 and 0.9241.

Sampling from Non Normal Population

- In many cases the data may not be normally distributed or the cases to be considered are very few to be approximated by normal distribution.

- The population distribution and the sampling distribution of the mean in such cases is different. The later tends to show normal distribution.

- Ifthe sample size is large then the sampling distribution of the mean approaches normality regardless of the shape of population distribution.

Central Limit Theorem

- The probability distribution suggests two things.
  - 1. Themean of the sampling distribution of the mean will equal the population meanregardless of the sample size even if the population is not normal.
  - 2. as the sample size increases the sampling distribution of the mean will approach normalcy, regardless of the shape of the population distribution.

- The above relationship between the shape of population distribution and the shape of the sampling distribution of the mean is called theCentral LimitTheorem.

- It assures that the sampling distribution of the mean approaches normalcy as the sample size increases.

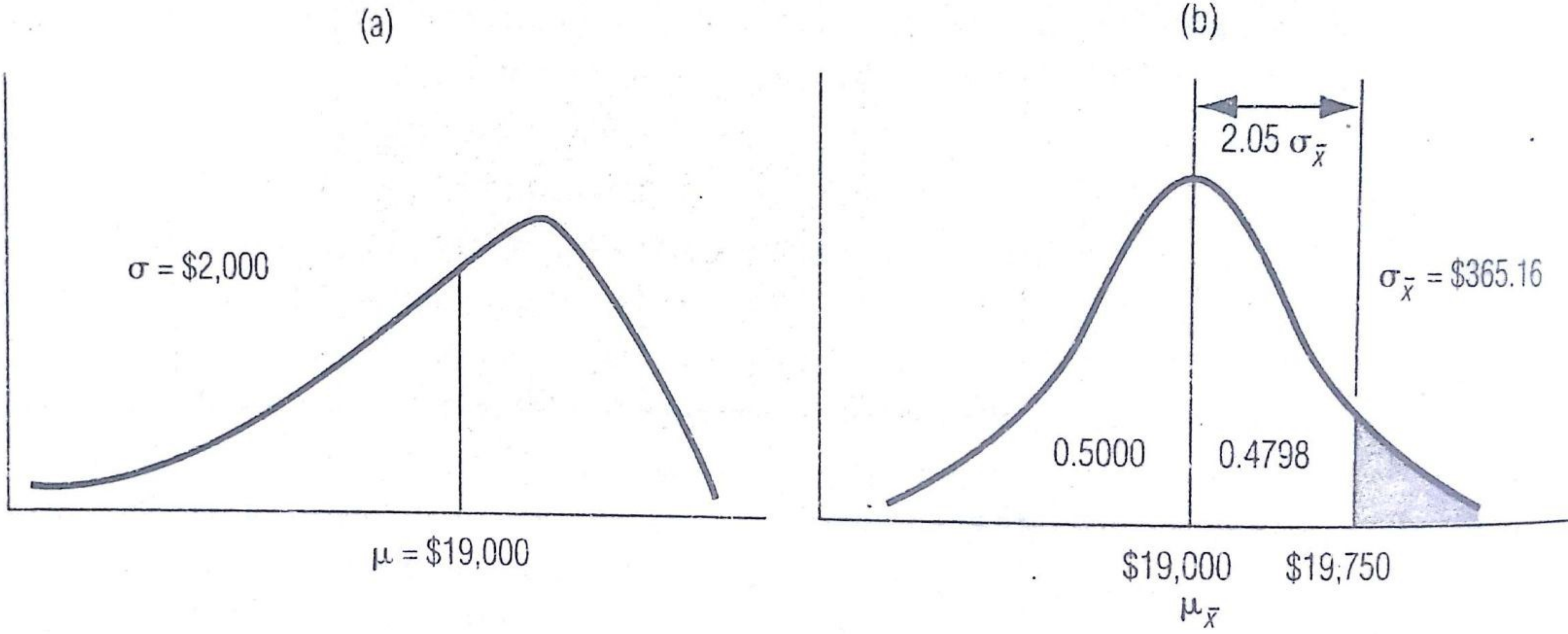
Advantages of CLM

- It assures that the sampling distribution of the mean approaches normalcy as the sample size increases.

- The significance of cml is that it permits us to use sample statistics to make differences about population parameters without knowing anything about the shape of the frequency distribution of tat population other tan what we can get from the sample.

Example

- The distribution of annual earnings of all bank tellers withfive years experience isskewed negatively. This distribution has a mean of $ 15000 and SD of $2000. If we draw the random sample of 30 tellers what is the probability that their earnings will average more than $15,750?
- $SE = \sigma \div \sqrt{n} = 365.16$
- $Z = 2.05$
- For the z value of 2.05 area is 0.4798. The area between the right hand tail and assumed average is 0.0202.
- There is slightly more than 2% chance of average earnings being more than $ 15,750.


(a)
σ = $2,000
μ = $19,000
(b)
2.05 σ_x̄
σ_x̄ = $365.16
0.5000
0.4798
$19,000
μ_x̄
$19,750

Problems to solve

- In a normal distribution with mean 56 and standard deviation 21 how large a sample must be taken so that there will be at least 90 percent chance that the mean is greater than 52?

- In a normal distribution with mean 375 and standard deviation 48 how large a sample must be taken so that the probability will be at least 0.95 that the sample mean falls between 370 and 380?

Relationship betweensample size and Standa error

- From the formula we notice that as n increases SE decreases.
- When n=10 SE= 31.63
- When n=100 then SE=10
- Thus by increasing the sample size tenfold the standard error dropped from 31.63 to 10, one third of the original value.
- Owing to the fact that SE varies inversely with the square root of n, there is a diminishing return in sampling.

Standard error of the mean for finite populat

- Many of the decision makers use finite population. In this case the formula for standard erroris $SE=\sigma/n*\surd N-n/N-1$
- Where N is the size of the population and

n is the size of the sample.

- The second term in the above formula is called as the finite population multiplier. Whenpopulationis very large in relation to the size of the sample, the finite population multiplier is close to 1.

Example

- There are 20 textile mills experiencingexcessive labour turnover. The study indicates that SD of the distribution of annual turnover is 75 employees. If 5 mills are chosen then

- $SE= SE=\sigma/\sqrt{n}*\sqrt{N-n/N-1}$
  - $=75/\ \sqrt{5} * \sqrt{20-5/20-1}$
  - $= 33.54* 0.888$
  - $= 29.8$
  - In this case the finite population multiplier of 0.888 reduced the standard error from 35.54 to 29.8.

Problems to solve

- From a population of 125 items with mean of 105 and standard deviation of 17, 64 items were chosen.
- A) what is the standard error
- B) what is the probability the value to fall between 107.5 and 109?

- From a population of 75 items with mean of 364 and variance of 18, 32 items were randomly selected without replacement.
- A) what s the standard error
- B) What is the probability for a value between 363 and 366.

Thank you,

Would monkey’s decision depend on sampling distributio