

Lecture 5a: Inferential statistics

Ernesto F. L. Amaral

October 28, 2024

Introduction to Sociological Data Analysis (SOCL 600)

www.ernestoamaral.com

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapter 6 (pp. 144–159).



Outline

- Explain the purpose of inferential statistics in terms of generalizing from a sample to a population
- Define and explain the basic techniques of random sampling
- Explain and define these key terms: population, sample, parameter, statistic, representative, EPSEM sampling techniques
- Differentiate between the sampling distribution, the sample, and the population
- Explain two theorems



Basic logic and terminology

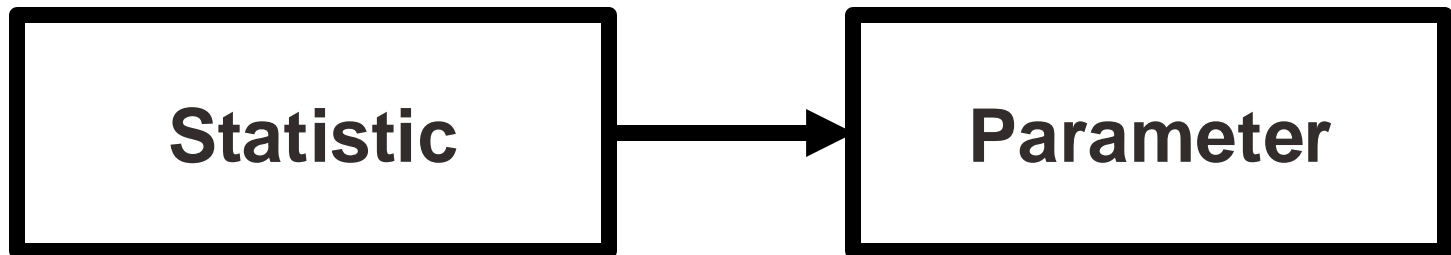
- **Problem**
- The populations we wish to study are almost always so large that we are unable to gather information from every case

- **Solution**
- We choose a sample – a carefully chosen subset of the population – and use information gathered from the cases in the sample to generalize to the population



Basic logic and terminology

- **Statistics** are mathematical characteristics of samples
- **Parameters** are mathematical characteristics of populations
- **Statistics** are used to estimate **parameters**



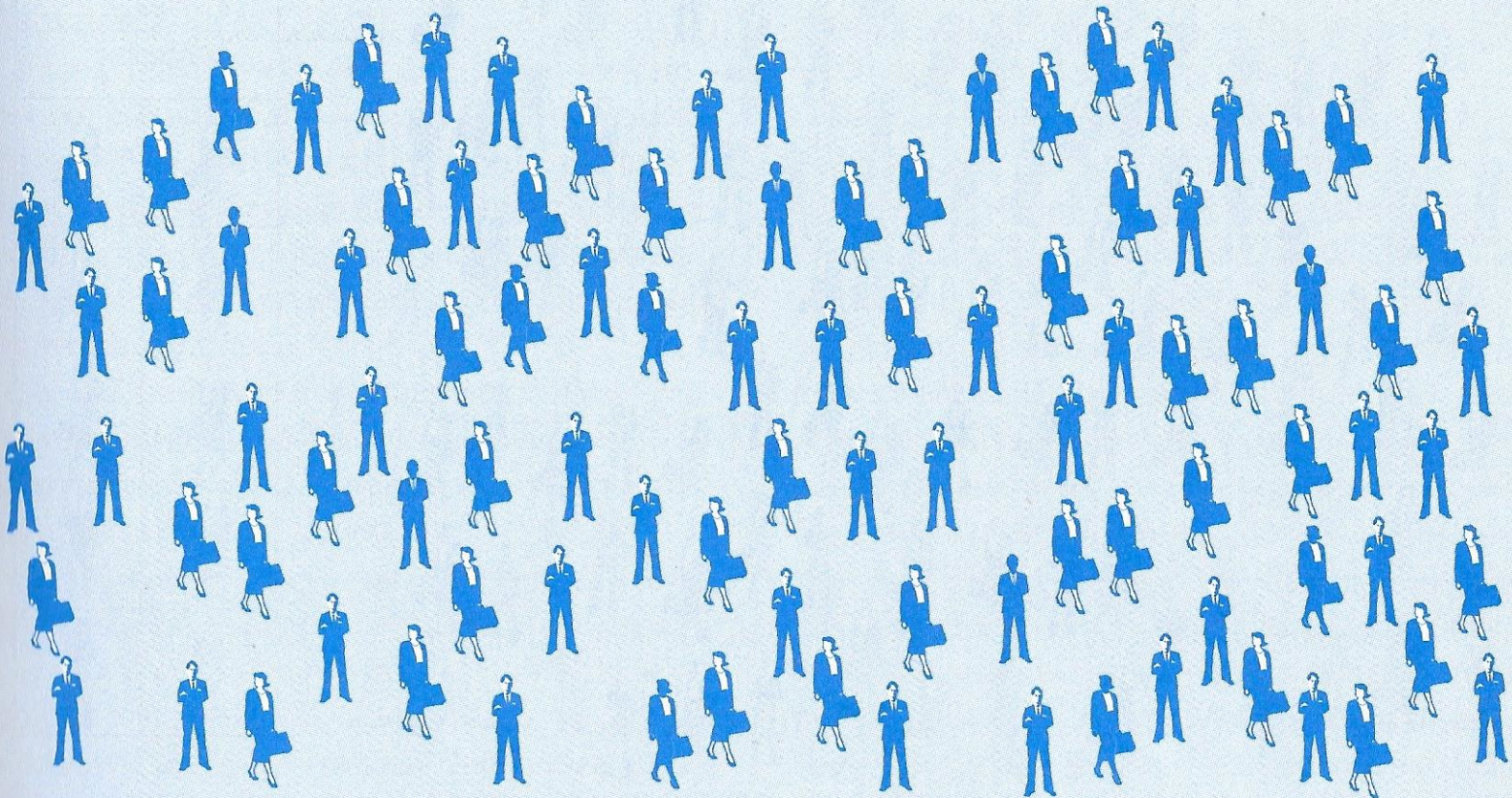
Samples

- Must be representative of the population
 - Representative: The sample has the same characteristics as the population
- How can we ensure samples are representative?
 - Samples drawn according to the rule of **EPSEM** (equal probability of selection method)
 - If every case in the population has the same chance of being selected, the sample is likely to be representative

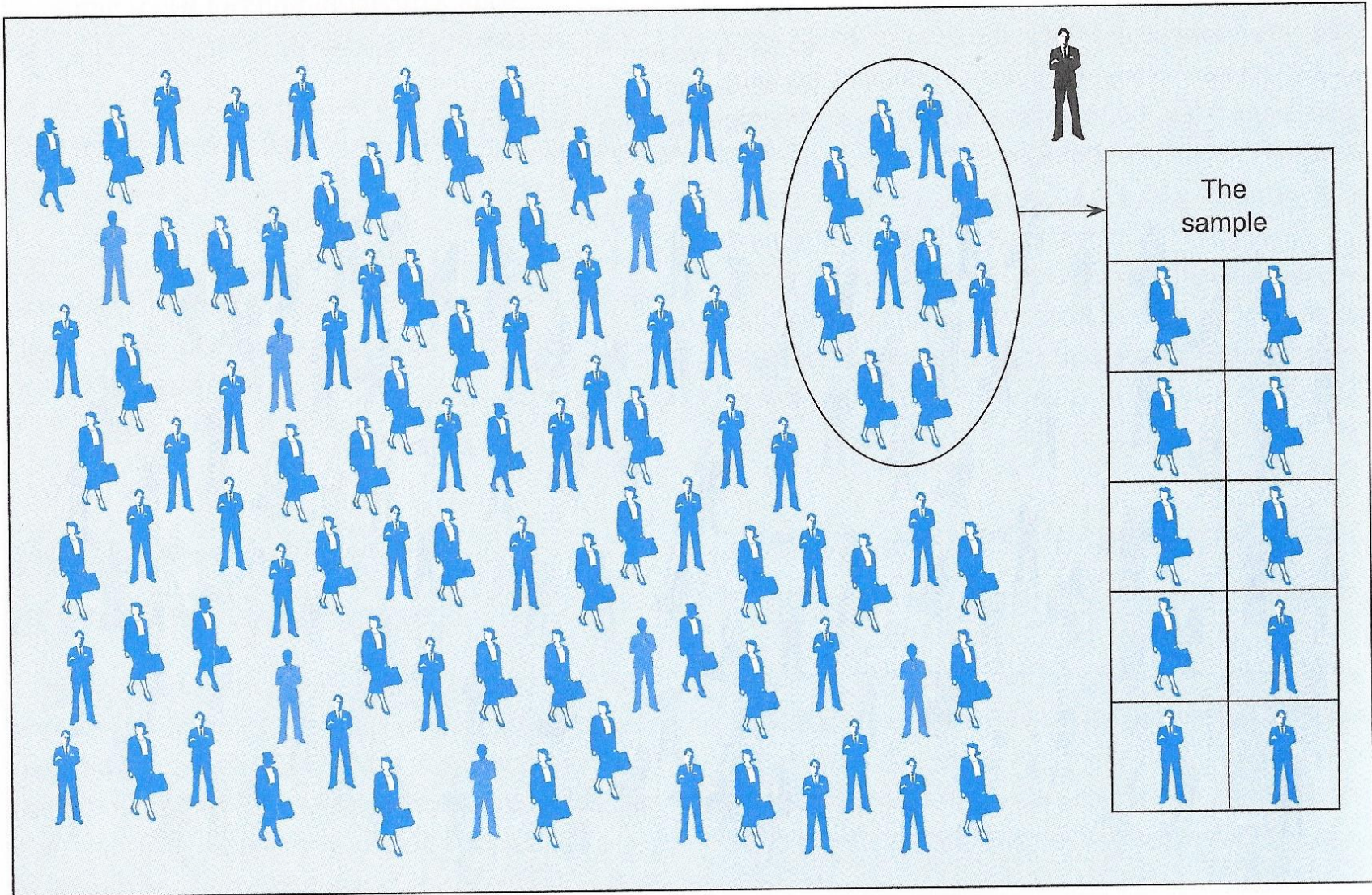


A population of 100 people

44 white women
44 white men
6 African-American women
6 African-American men



Nonprobability sampling



EPSEM sampling techniques

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling



1. Simple random sampling

- To begin, we need
 - A list of the population
- Then, we need a method for selecting cases from the population, so each case has the same probability of being selected
 - The principle of EPSEM
 - A sample selected this way is very likely to be representative of the population
 - Variable in population should have a normal distribution or $n > 30$



Example

- You want to know what percent of students at a large university work during the semester
- Draw a sample size (n) of 500 from a list of all students ($N=20,000$)
- Assume the list is available from the Registrar
- How can you draw names, so every student has the same chance of being selected?



Example

- Each student has a unique, 6 digit ID number that ranges from 000001 to 999999
- Use a table of random numbers or a computer program to select 500 ID numbers with 6 digits each
- Each time a randomly selected 6 digit number matches the ID of a student, that student is selected for the sample
- Continue until 500 names are selected



Example

- **Stata**

```
set obs 500
```

```
generate student = runiformint(1,999999)
```

```
sum student
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
student	500	482562.6	283480.9	3652	997200

- **Excel**

- **RANDBETWEEN** (minimum,maximum)

- Returns a random number between those you specify
- Drag the function to 500 cells

=RANDBETWEEN(1,999999)

- **RANDARRAY** (rows,columns,minimum,maximum)

=RANDARRAY(500,1,1,999999)



Example

- Disregard duplicate numbers
- Ignore cases in which no student ID matches the randomly selected number
- After questioning each of these 500 students, you find that 368 (74%) work during the semester

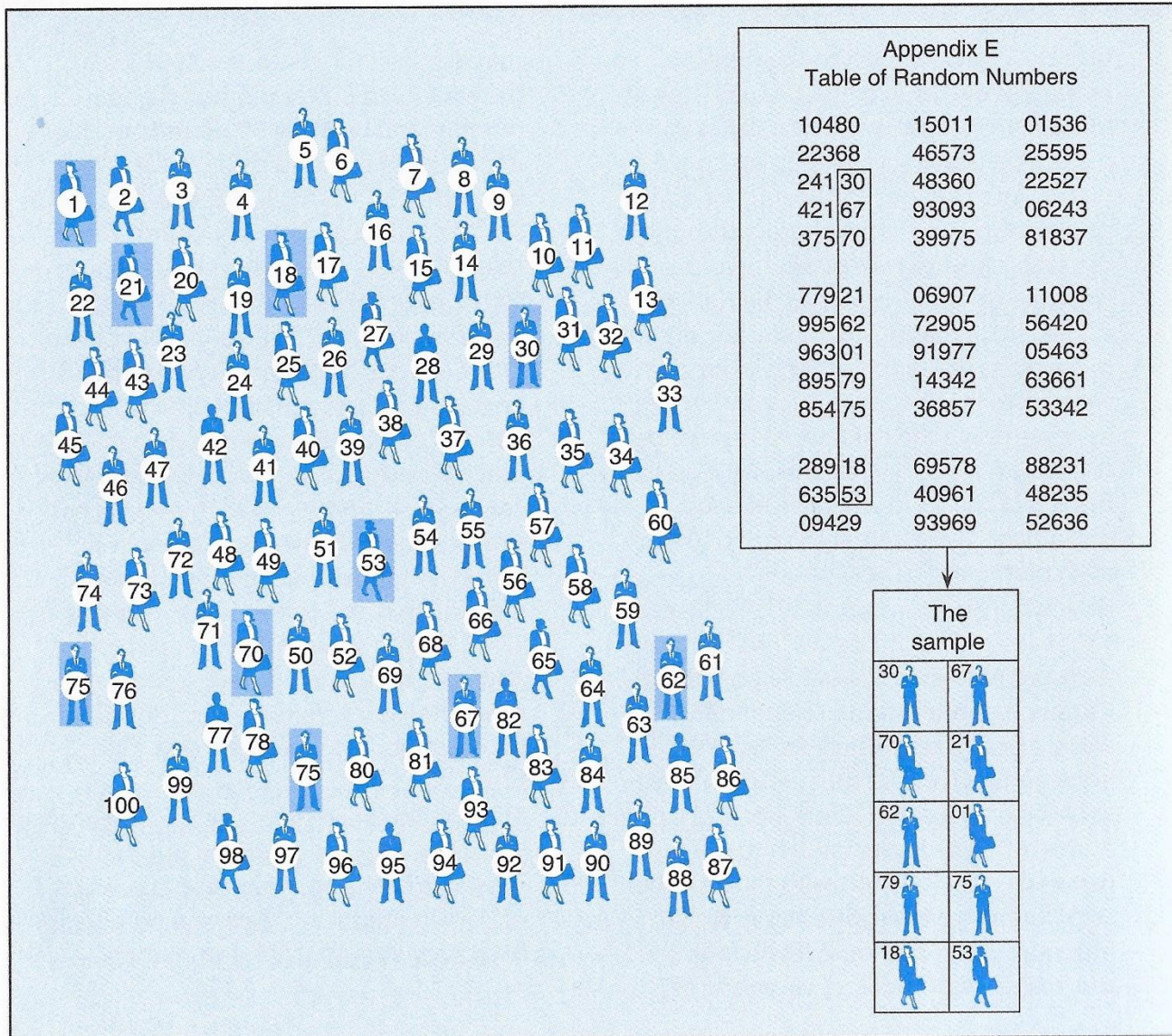


Applying logic and terminology

- In the previous example:
- **Population:** All 20,000 students
- **Sample:** 500 students selected and interviewed
- **Statistic:** 74% (percentage of sample that held a job during the semester)
- **Parameter:** Percentage of all students in the population who held a job



Simple random sample



2. Systematic sampling

- Useful for large populations
- Randomly select the first case then select every k^{th} case
- **Sampling interval**
 - Distance between elements selected in the sample
 - Population size (N) divided by sample size (n)
- **Sampling ratio**
 - Proportion of selected elements in the population
 - Sample size (n) divided by population size (N)
- Can be problematic if the list of cases is not truly random or demonstrates some patterning



Example

- If a list contained 10,000 elements and we want a sample of 1,000
- Sampling interval
 - Population size / sample size = $10,000 / 1,000 = 10$
 - We would select every 10th element for our sample
- Sampling ratio
 - Sample size / population size = $1,000 / 10,000 = 1/10$
 - Proportion of selected elements in population
- Select the first element at random



3. Stratified sampling

- It guarantees the sample will be representative on the selected (stratifying) variables
 - Stratification variables relate to research interests
- First, divide the population list into subsets, according to some relevant variable
 - **Homogeneity within subsets**
 - E.g., only women in a subset; only men in another subset
 - **Heterogeneity between subsets**
 - E.g., subset of women is different than subset of men
- Second, sample from the subsets
 - Select the number of cases from each subset proportional to the population

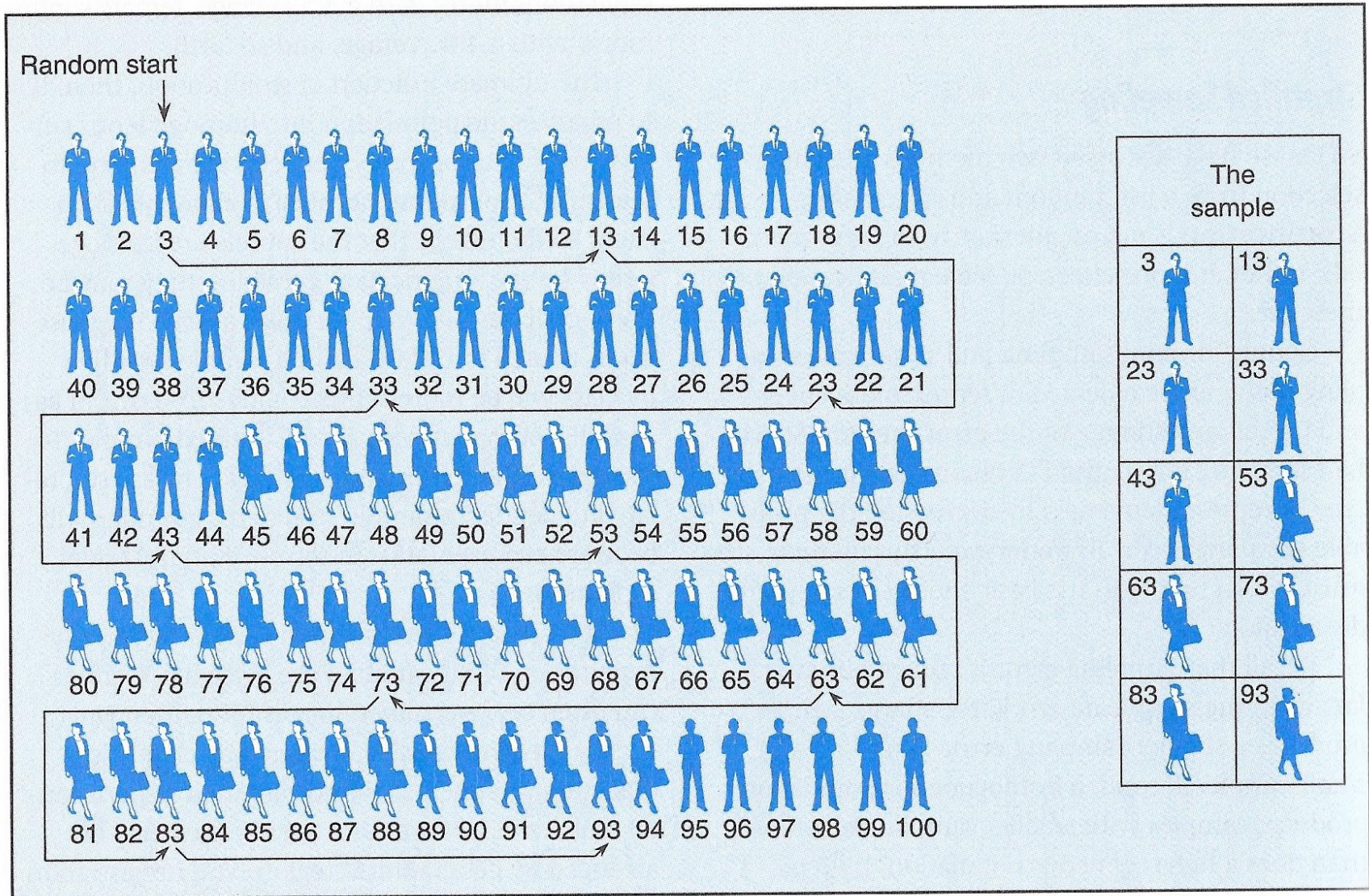


Example

- If you want a sample of 1,000 students
 - That would be representative to the population of students by sex and GPA
- You need to know the population composition
 - E.g., women with a 4.0 average compose 15 percent of the student population
- Your sample should follow that composition
 - In a sample of 1,000 students, you would select 150 women with a 4.0 average



Stratified, systematic sample



4. Cluster sampling

- Select groups (or clusters) of cases rather than single cases
 - **Heterogeneity within subsets**
 - E.g., each subset has both women and men, following same proportional distribution as population
 - **Homogeneity between subsets**
 - E.g., all subsets with both women and men should be similar
- Clusters are often geographically based
 - For example, cities or voting districts
- Sampling often proceeds in stages
 - Multi-stage cluster sampling
 - Less representative than simple random sampling



Stratified vs. cluster sampling

- **Stratified**

- Homogeneity within subsets
- Heterogeneity between subsets
- Select cases from each subset

Subset of
women

Subset of
men

- **Cluster**

- Heterogeneity within subsets (groups, clusters, areas)
- Homogeneity between subsets
- Select groups (e.g., area 1) rather than single cases

Area 1:
women & men

Area 2:
women & men



Sampling distribution

- Sampling distribution is the probabilistic distribution of a statistic for all possible samples of a given size (n)
 - It is the distribution of a statistic (e.g., proportion, mean) for all possible outcomes of a certain size
- Central tendency and dispersion
 - Mean is the same as the population mean
 - Standard deviation is referred as standard error
 - It is the population standard deviation divided by the square root of n
 - We have to take into account the complex survey design to estimate the standard error (`svyset` command in Stata)



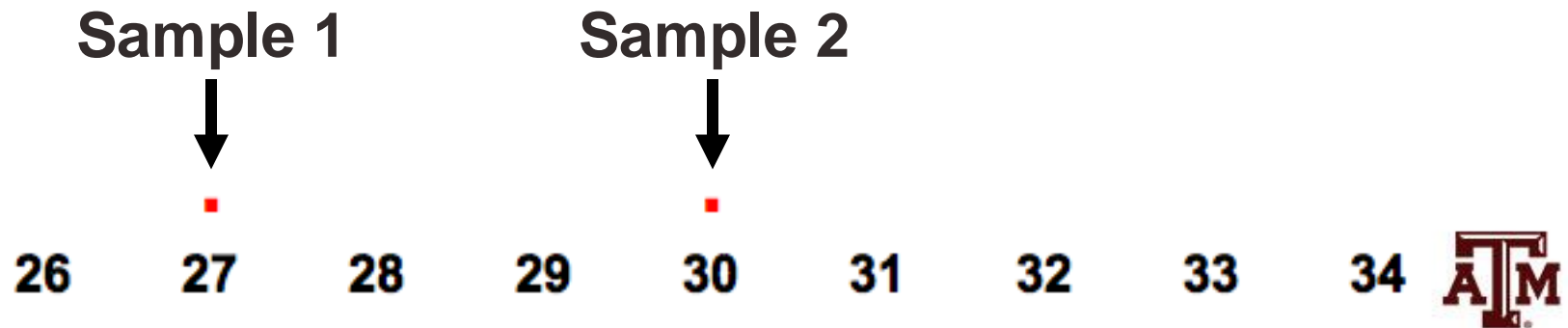
Linking sample and population

- Every application of inferential statistics involves three different distributions
 - Population: empirical; unknown
 - Sampling distribution: theoretical; known
 - Sample: empirical; known
- In inferential statistics, the sample distribution links the sample with the population



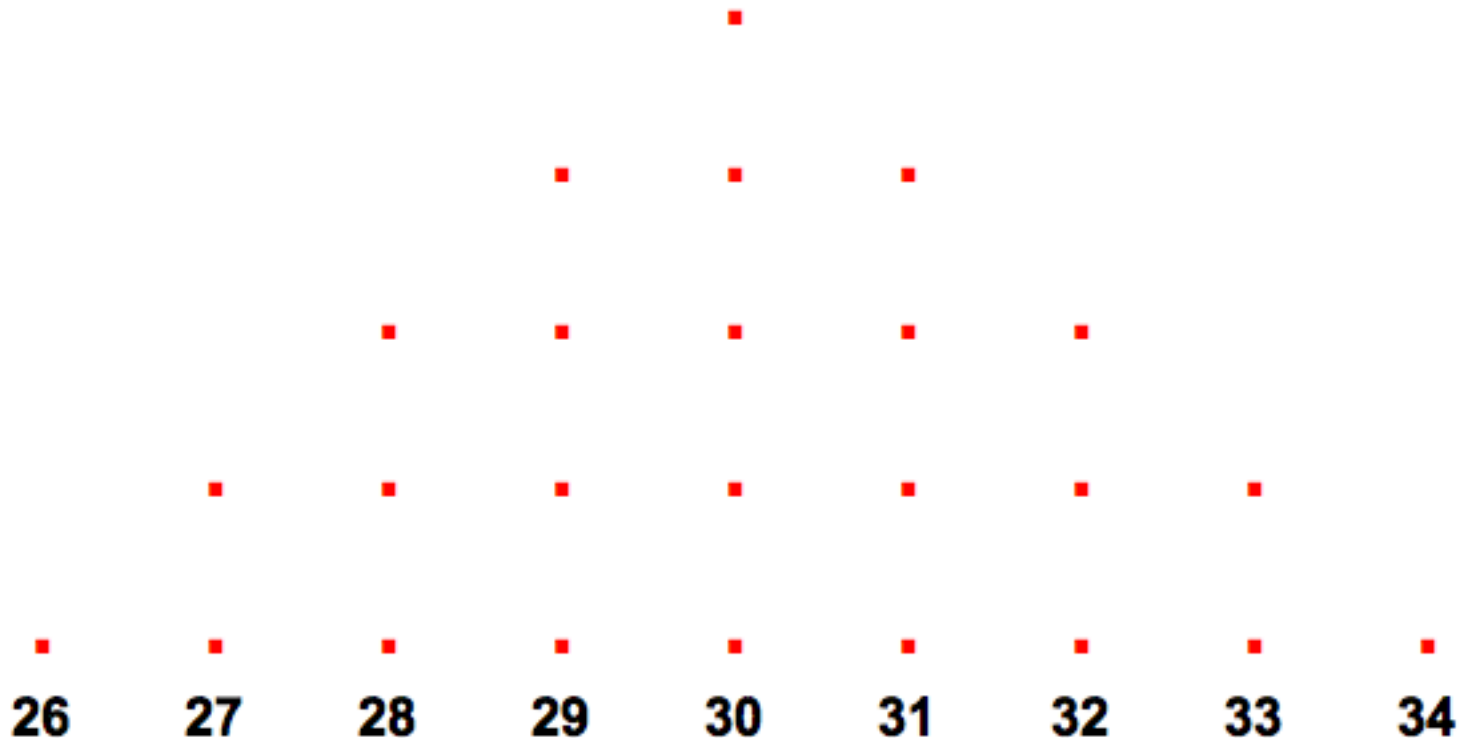
Example

- Suppose we want to gather information on the age of a community of 10,000 individuals
 - Sample 1: $n=100$ people, plot sample's mean of 27
 - Replace people in the sample back to the population
 - Sample 2: $n=100$ people, plot sample's mean of 30
 - Replace people in the sample back to the population

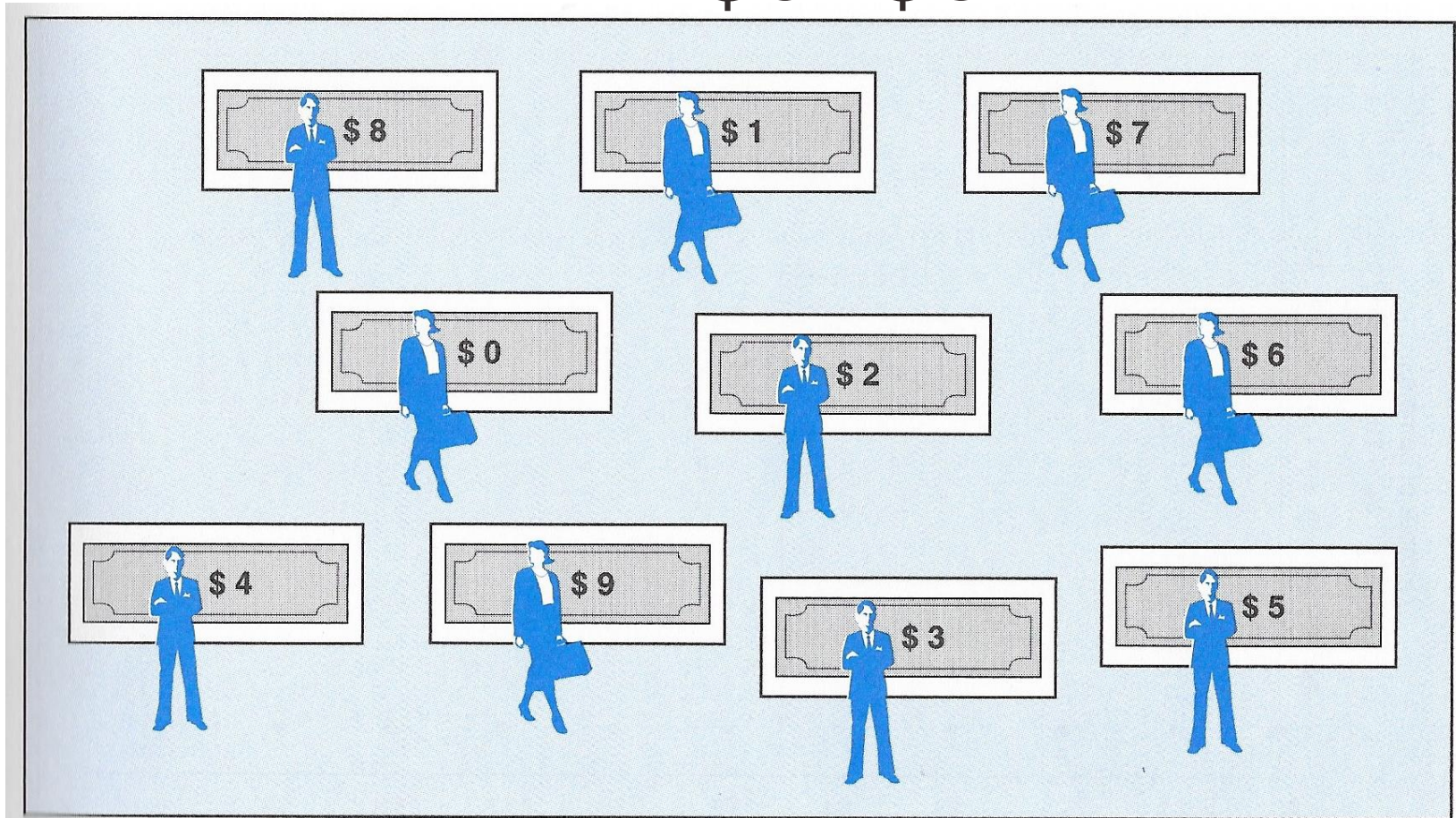


Example

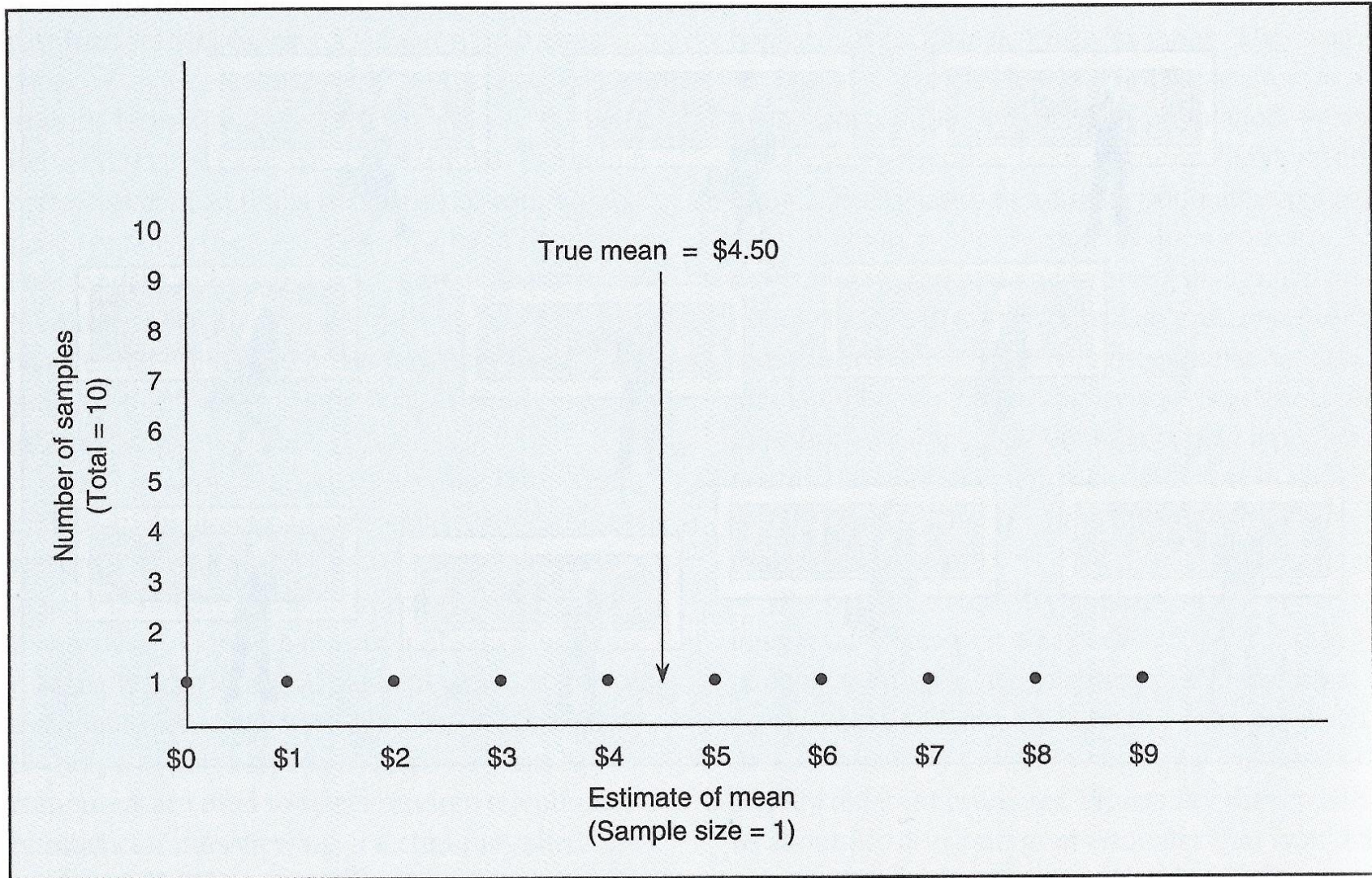
- We repeat this procedure: sampling, replacing
 - Until we have exhausted every possible combination of 100 people from the population of 10,000
 - Sampling distribution has a normal shape



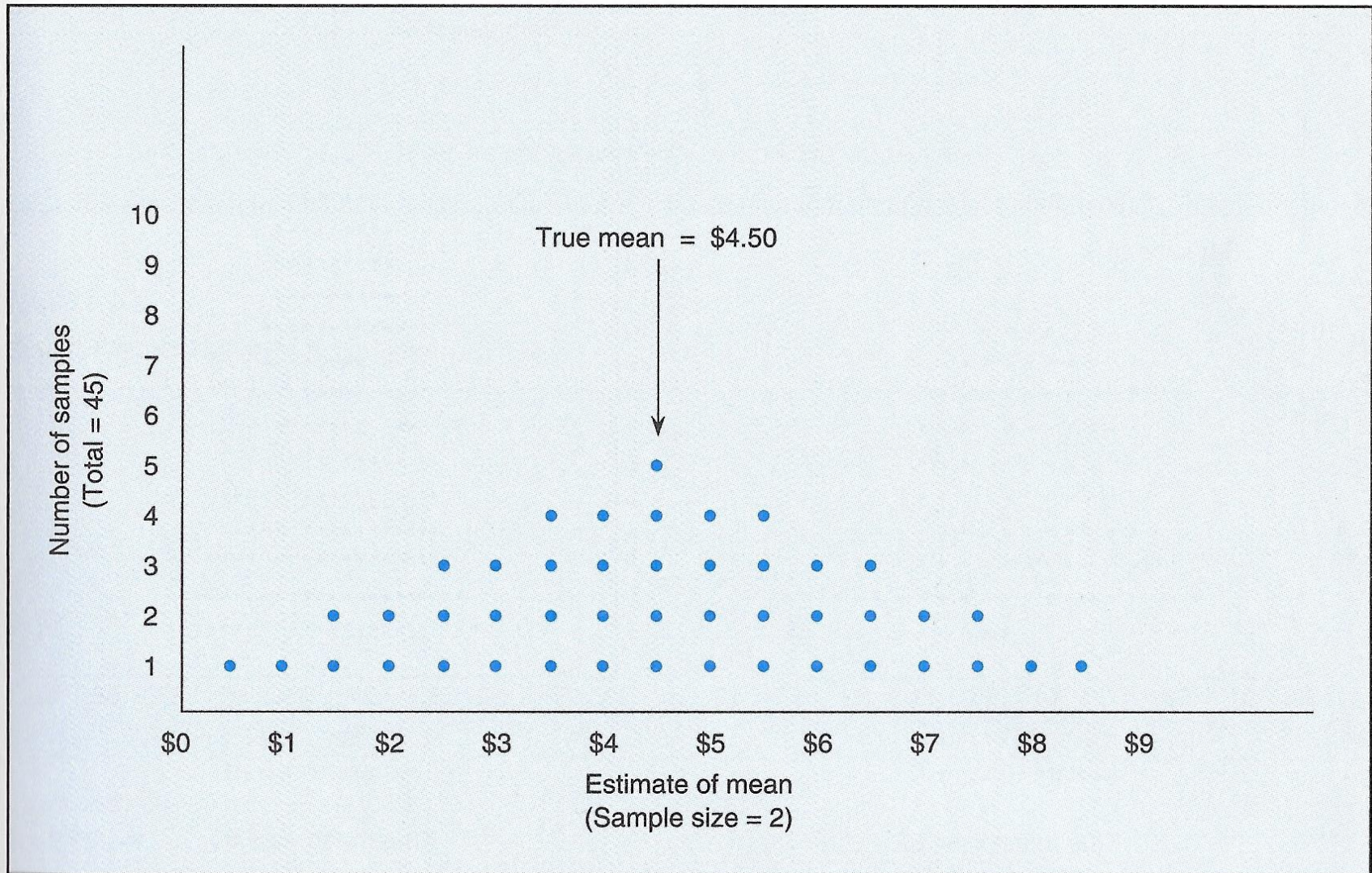
Another example: A population of 10 people with \$0–\$9



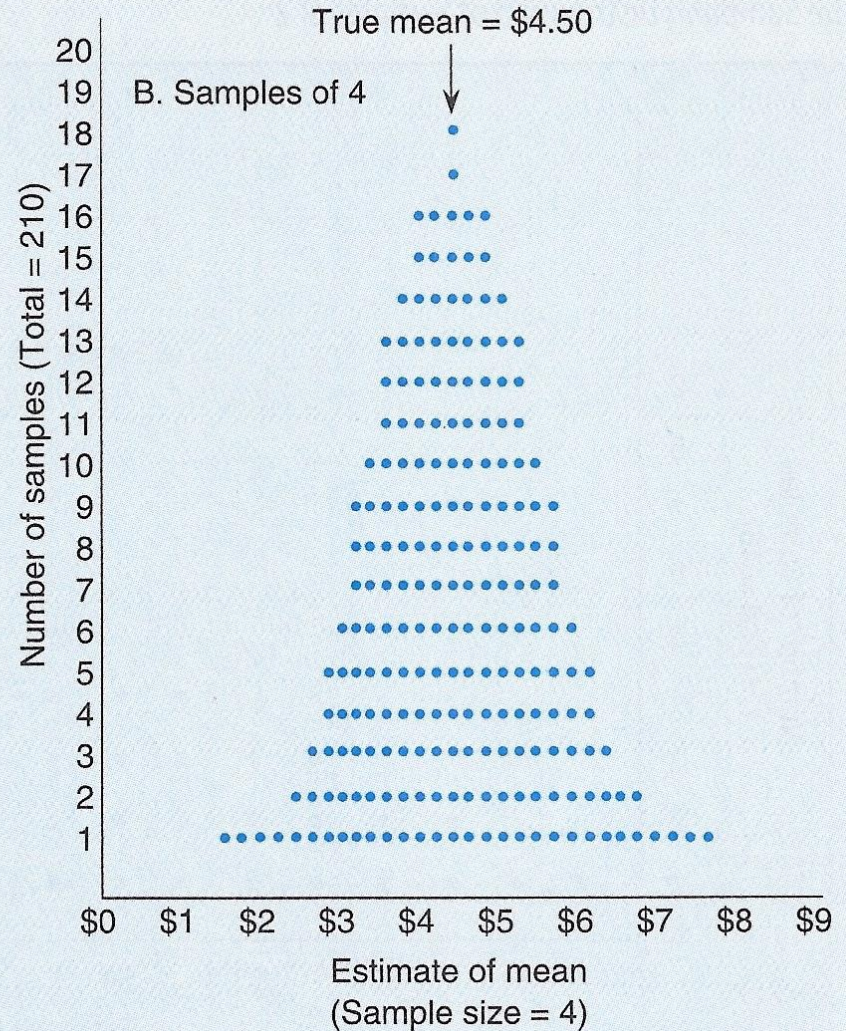
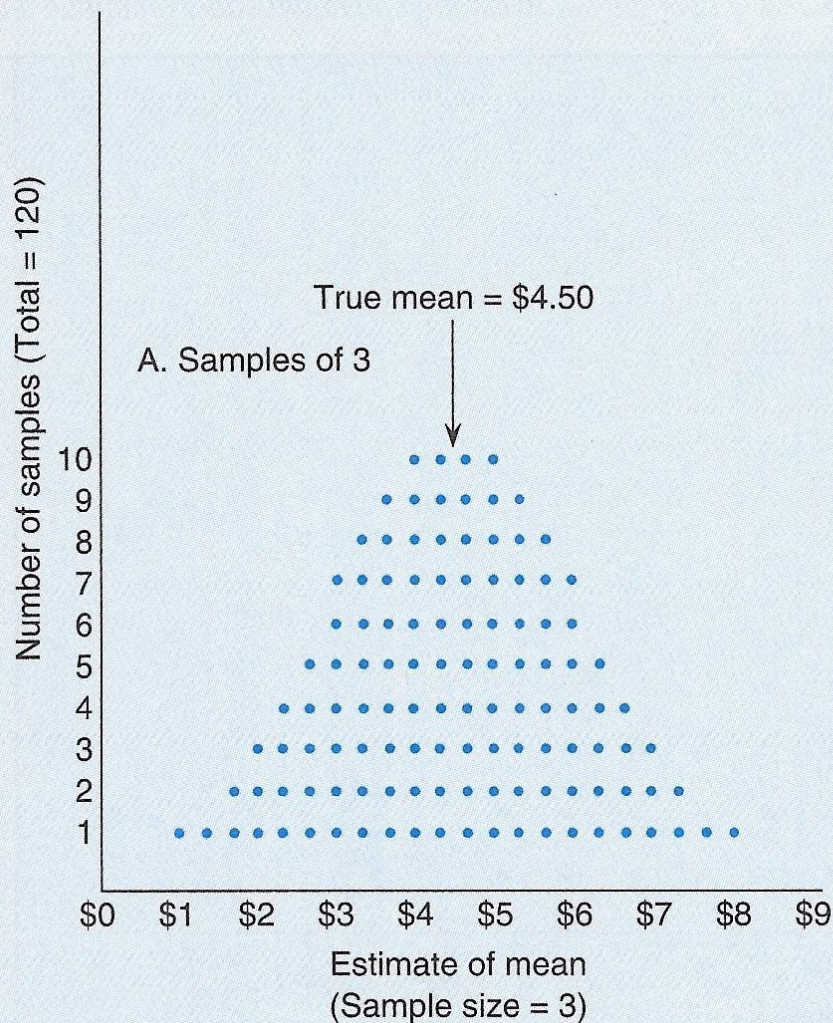
The sampling distribution ($n=1$)



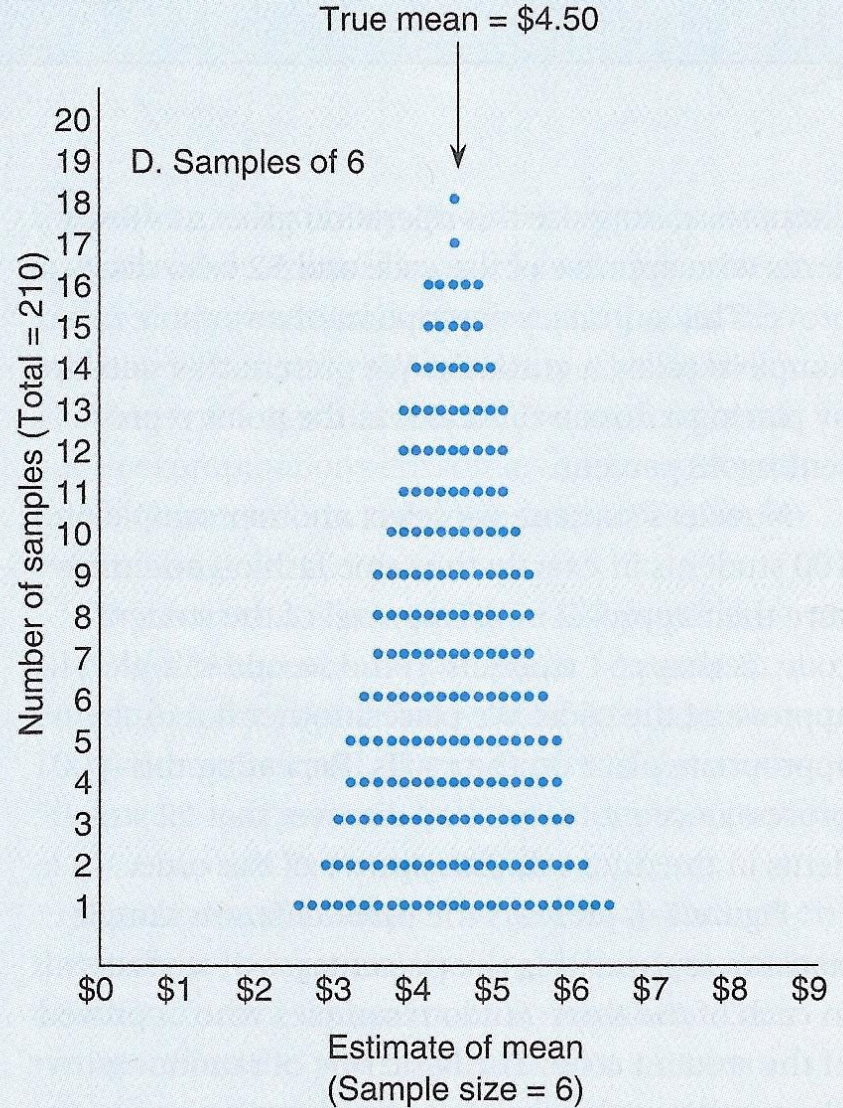
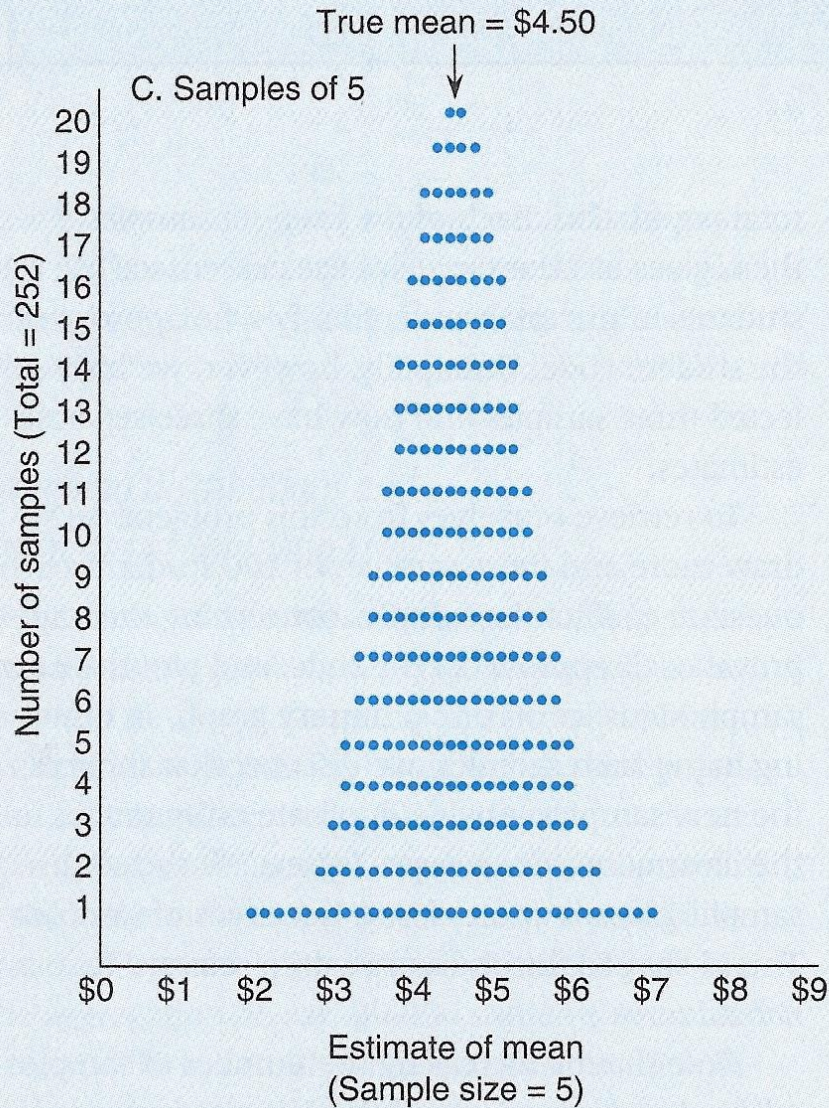
The sampling distribution ($n=2$)



The sampling distribution



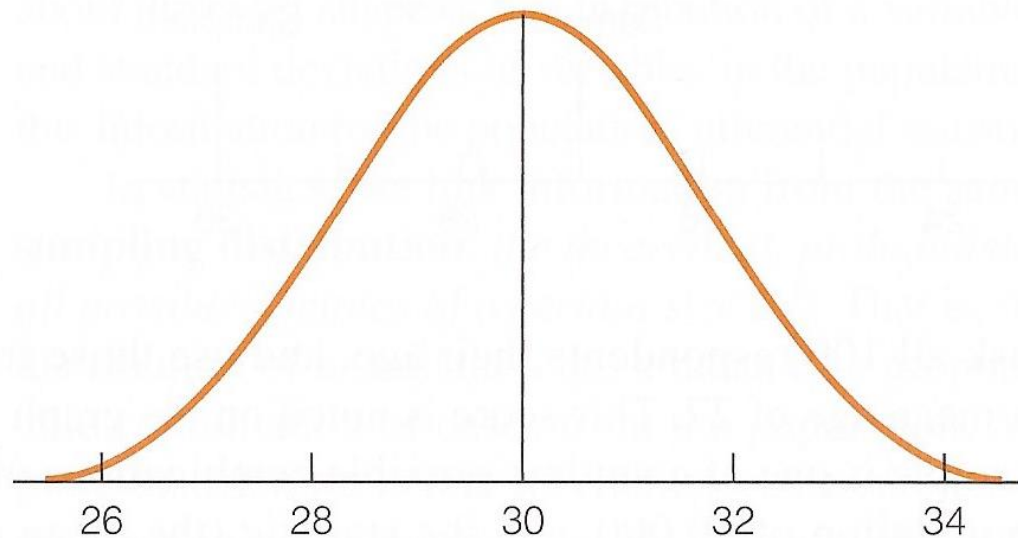
The sampling distribution



Properties of sampling distribution

- It has a mean ($\mu_{\bar{X}}$) equal to the population mean (μ)
- It has a standard deviation (standard error, $\sigma_{\bar{X}}$) equal to the population standard deviation (σ) divided by the square root of n
- It has a normal distribution

A Sampling Distribution of Sample Means



First theorem

- Tells us the shape of the sampling distribution and defines its mean and standard deviation
- If repeated random samples of size n are drawn from a **normal population** with mean μ and standard deviation σ
 - Then, the sampling distribution of sample means will **have a normal distribution** with...
 - A mean: $\mu_{\bar{x}} = \mu$
 - A standard error of the mean: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$



First theorem

- Begin with a characteristic that is normally distributed across a population (IQ, height)
- Take an infinite number of equally sized random samples from that population
- The sampling distribution of sample means will be normal

Central limit theorem

- If repeated random samples of size n are drawn from **any population** with mean μ and standard deviation σ
 - Then, as n becomes large, the sampling distribution of sample means will **approach normality** with...
 - A mean: $\mu_{\bar{x}} = \mu$
 - A standard error of the mean: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
- This is true for any variable, even those that are not normally distributed in the population
 - As sample size grows larger, the sampling distribution of sample means will become normal in shape



Central limit theorem

- The importance of the central limit theorem is that it removes the constraint of normality in the population
 - Applies to large samples ($n \geq 100$)
- If the sample is small ($n < 100$)
 - We must have information on the normality of the population before we can assume the sampling distribution is normal

Additional considerations

- The sampling distribution is normal
 - We can estimate areas under the curve (Appendix A)
 - Or in Stata: `display normal(z)`
- We do not know the value of the population mean (μ)
 - But the mean of the sampling distribution ($\mu_{\bar{x}}$) is the same value as μ
- We do not know the value of the population standard deviation (σ)
 - But the standard deviation of the sampling distribution ($\sigma_{\bar{x}}$) is equal to σ divided by the square root of n



Symbols

Distribution	Shape	Mean	Standard deviation	Proportion
Samples	Varies	\bar{X}	s	P_s
Populations	Varies	μ	σ	P_u
Sampling distributions	Normal	$\mu_{\bar{X}}$		
of means		$\mu_{\bar{X}}$	$\sigma_{\bar{X}} = \sigma/\sqrt{n}$	
of proportions		μ_p	σ_p	





TEXAS A&M
UNIVERSITY.