# SNS COLLEGE OF TECHNOLOGY

**(An Autonomous Institution)**

Re-accredited by NAAC with A+ grade, Accredited by NBA(CSE, IT, ECE, EEE & Mechanical)

Approvedy by AICTE, New Delhi, Recognized by UGC, Affiliated to Anna University, Chennai

## Department of MCA

Topic: SAMPLING DISTRIBUTION

**Course**

**19CAT702**
**Big Data Analytics**

**Unit I**

**Introduction to Big data**

**Elective**

**II Semester / I MCA**

# Introduction to Sampling

**What is sample?**

• A sample is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population”

**Why sample?**

– Resources (time, money) and workload

– Gives results with known accuracy that can be calculated mathematically

• The sampling frame is the list from which the potential respondents are drawn

– Registrar’s office

– Class rosters

– Must assess sampling frame errors

# Introduction to Sampling

**Major Types of Samples**

**Probability (Random) Samples**

- Simple random sample
  - Systematic random sample
  - Stratified random sample
  - Multistage sample
  - Multiphase sample
  - Cluster sample
- **Non-Probability Samples**
  - Convenience sample
  - Purposive sample
  - Quota

# Specific Types of Samples

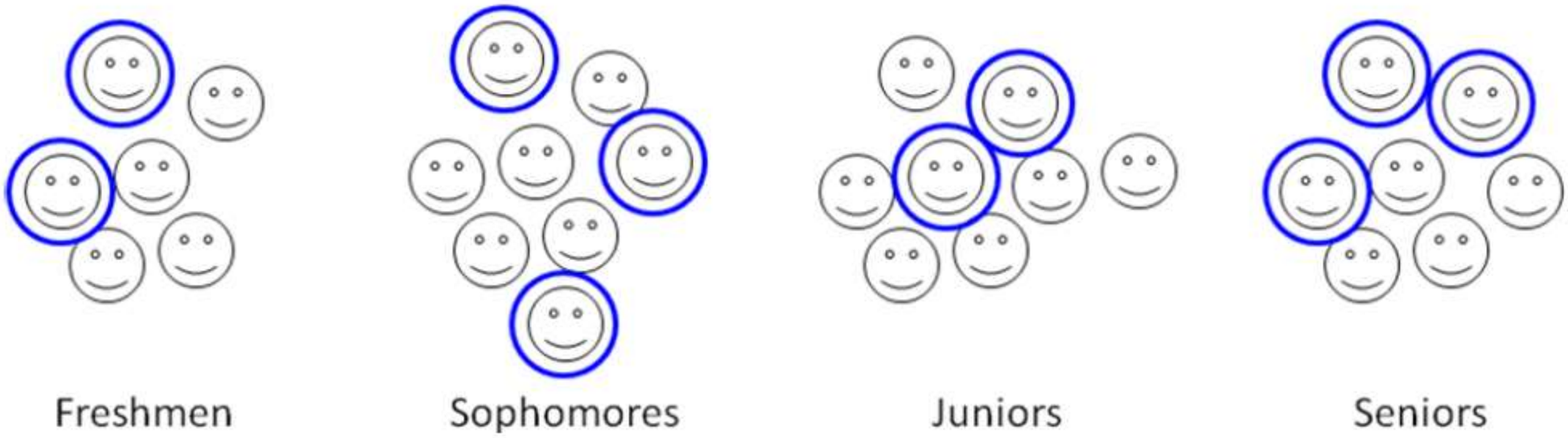
1. **Stratified Samples**
2. **Cluster Samples**
3. **Systematic Samples**
4. **Convenience Samples**

# 1. Stratified Samples

A **stratified sample has:**

i. members from each segment of a population.

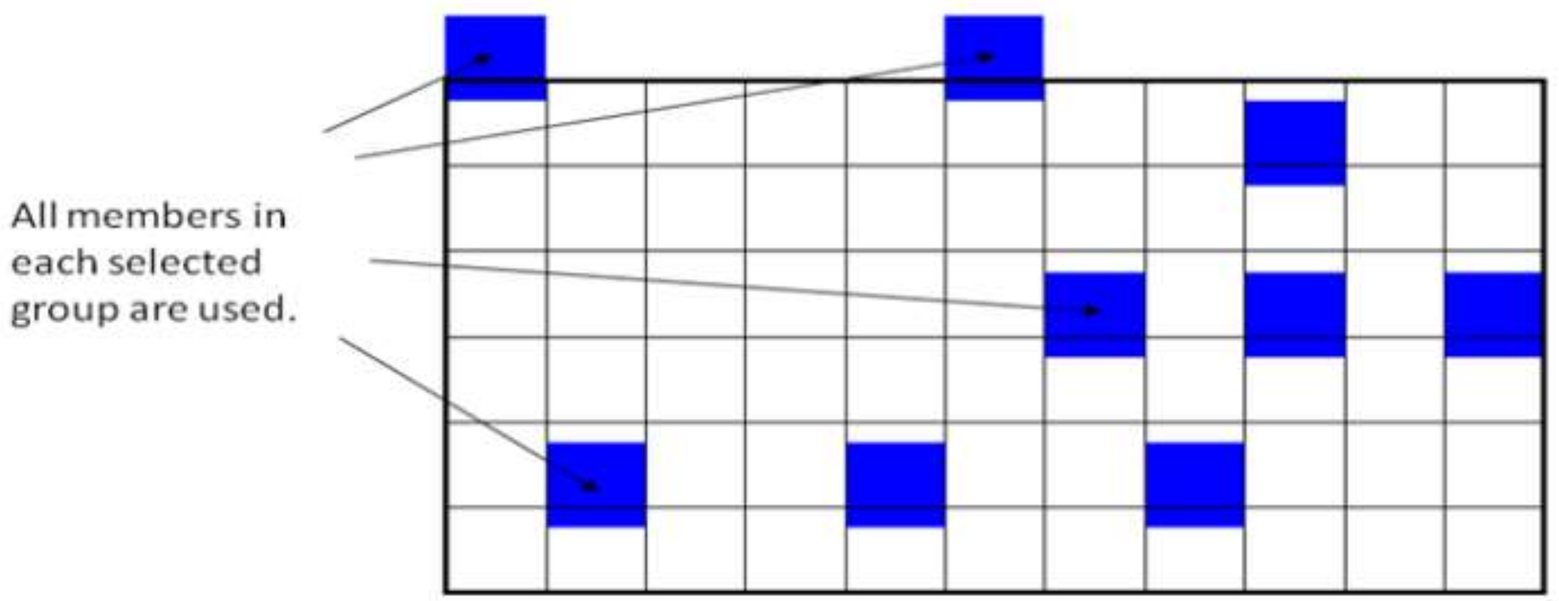
ii. This ensures that each segment from the population is represented.



# 2. Cluster Samples

A **cluster sample has**

- all members from randomly selected segments of a population.
- This is used when the population falls into naturally occurring subgroups.

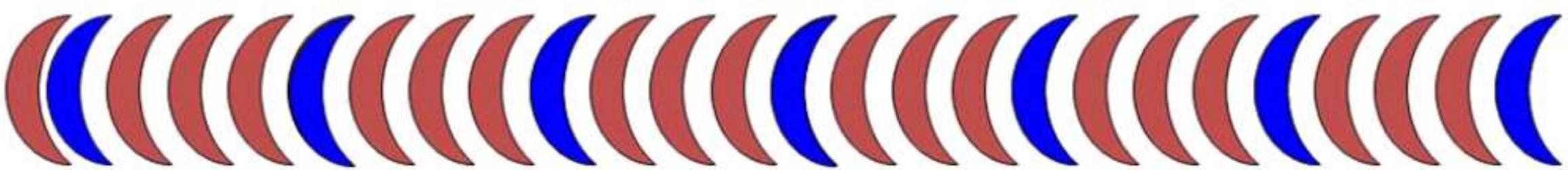



The city of Clarksville divided into city blocks.

# 3. Systematic Samples

A **systematic sample is**

- a sample **in which each member of the population is assigned a number.**
- A starting number is
- **randomly selected and sample members are selected at regular intervals.**



Every fourth member is chosen.

# 4. Convenience Samples

**A convenience sample consists only of available members of the population.**

**Example:**

- You are doing a study to determine the number of years of education each teacher at your college has.
- Identify the sampling technique used if you select the samples listed.

# Sampling Distribution

**What is Sampling Distribution?**

- **The way our means would be distributed**

– if we (1) ***collected a sample,***
– recorded the mean and
– threw it back, and
– (2) ***collected another,***
– recorded the mean and
– threw it back, and
– did this again and again….

# Sampling Distribution

- From Vogt:

A theoretical frequency distribution of the scores for or values of a statistic, such as a mean.

– Any statistic that can be computed for a sample has a sampling distribution.

- A sampling distribution is:

– the distribution of statistics

– that *would be produced*

– in repeated random sampling (with replacement) from the same population.

# Glimpses of Sampling Distribution

- **Sampling distributions is**
  - *all possible values of a statistic and*
  - *their probabilities of occurring for a sample of a particular size.*
- **Sampling distributions are used to**
  - calculate the probability that sample statistics
  - could have occurred by chance and
  - thus to decide whether something that is true of a sample statistic is
- also likely to be true of a population parameter.

# A Positive move of Sampling Distribution

- We are moving **from descriptive statistics to inferential statistics.**
- Inferential statistics allow the researcher:

– **to come to conclusions about a population**

– **on the basis of descriptive statistics about a sample.**

**Examples of Sampling Distribution**

1) Your sample says that

**a candidate gets support from 47%.**

**2) Inferential statistics allow you to say that**

– (a) the candidate gets support from 47% of the population

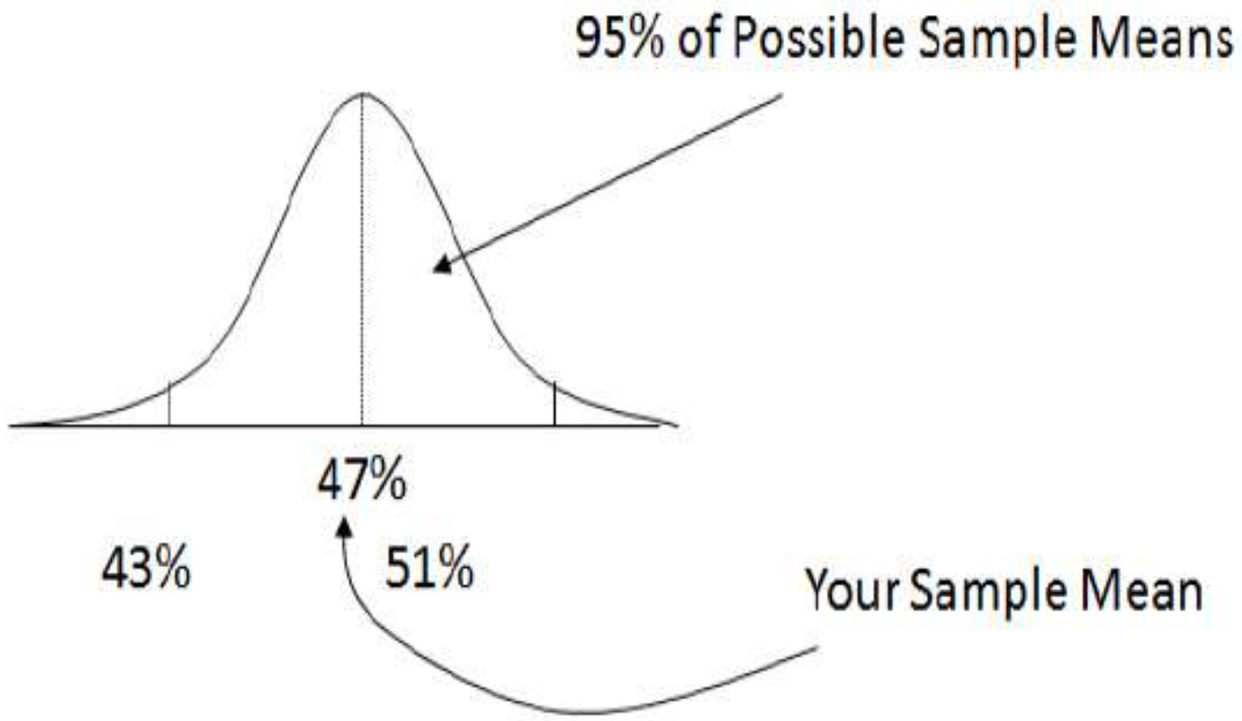
– (b) with a margin of error of +/- 4%.

– This means that the support in the population is

**likely somewhere between 43% and 51%.**

# Error on Sampling Distribution

- Margin of error is taken directly from a sampling distribution.
- It looks like this:



# A Real Time Scenario on Sampling Distribution

Let’s create a sampling distribution of means…
a) Take a sample of size 1,500 from the US.
b) Record the mean income.
c) Our census said the mean is $30K.
Let’s create a sampling distribution of means…

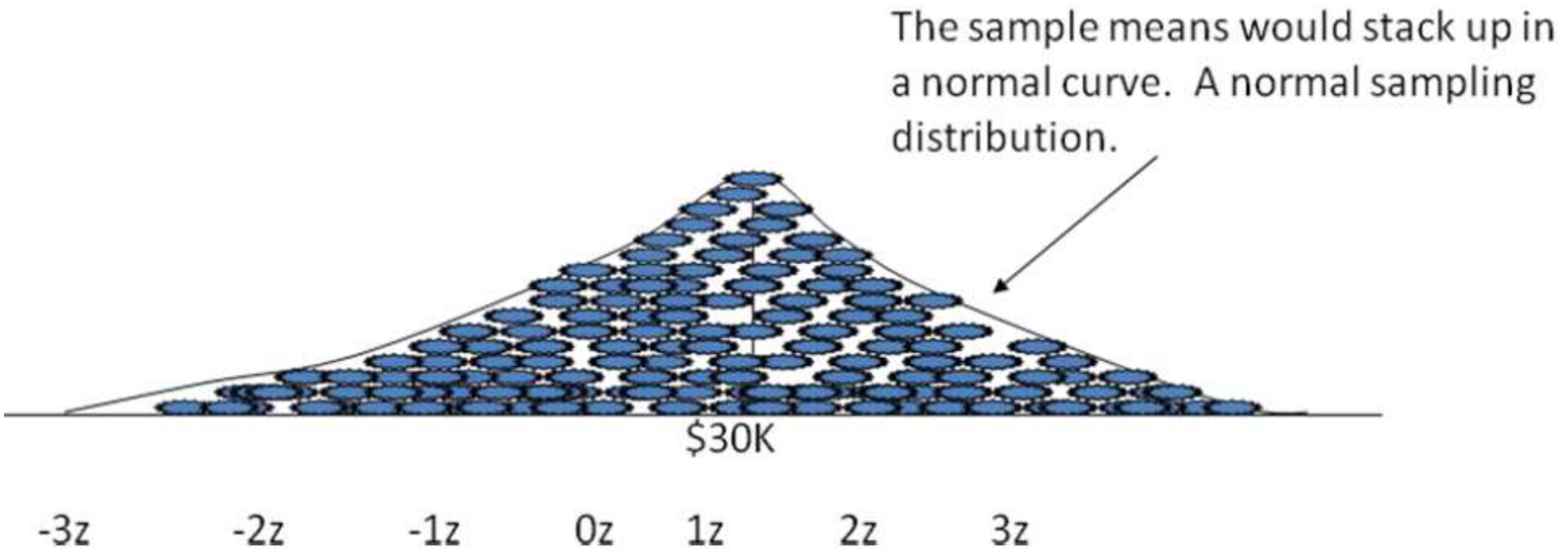
Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.

Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.

Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.

# A Real Time Scenario on Sampling Distribution

Say that the standard deviation of this distribution is $10K.
Think back to the empirical rule. What are the odds you would get a sample mean that is more than $20K off.



# A Real Time Scenario on Sampling Distribution

Knowing the likely variability of the sample means from repeated sampling gives us a context within which to judge how much we can trust the number we got from our sample.

For example, if the variability is low, , we can trust our number more than if the variability is high

• The first sampling distribution above, a, has a lower standard error.

Now a definition!

The standard deviation of a normal sampling distribution is called the standard error.

# A Real Time Scenario on Sampling Distribution

- **Statisticians have found that**

– the *standard error of a sampling distribution is :*

- **quite directly affected by**
- the number of cases in the sample(s), and
- the variability of the population distribution.

Population Variability:

For example, Americans' incomes are quite widely distributed, from $0 to Bill Gates'.

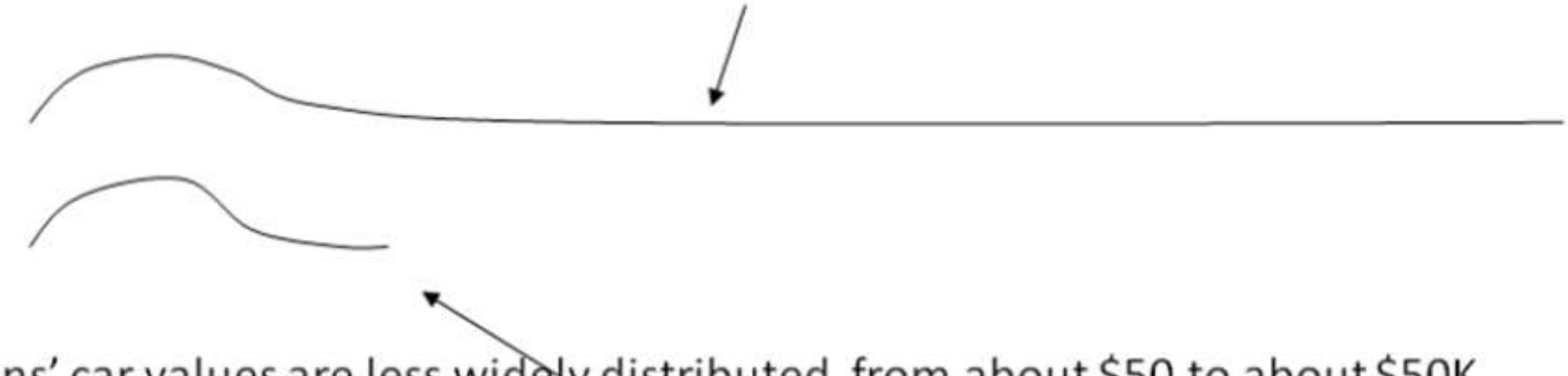
Americans' car values are less widely distributed, from about $50 to about $50K.

The standard error of the latter's sampling distribution will be a lot less variable.

# A Real Time Scenario on Sampling Distribution

Population Variability:

For example, Americans' incomes are quite widely distributed, from $0 to Bill Gates'.

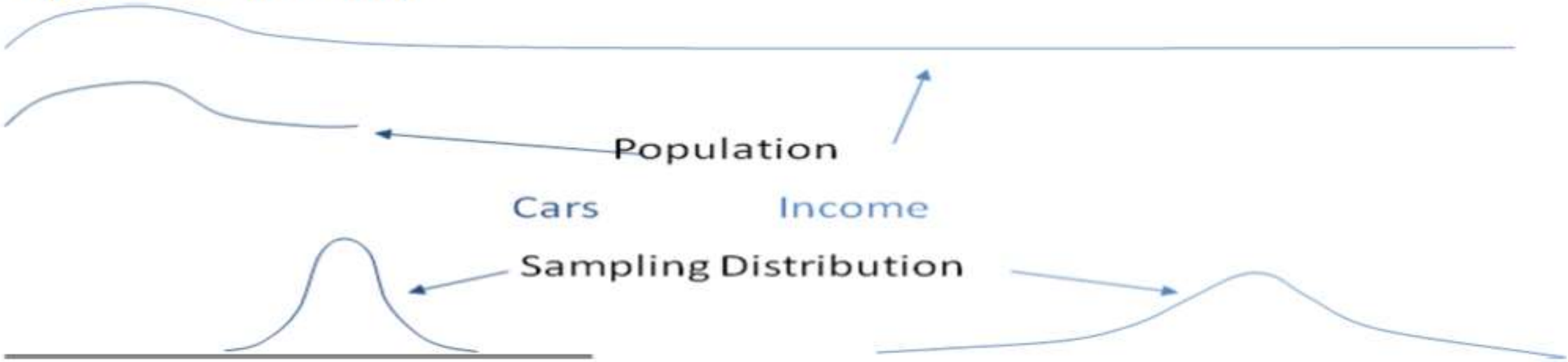


Americans' car values are less widely distributed, from about $50 to about $50K.

The standard error of the latter's sampling distribution will be a lot less variable.

# A Real Time Scenario on Sampling Distribution

**Population Variability:**




The standard error of income's sampling distribution will be a lot higher than car price's.

# What decides the Sampling Distribution?

Standard error :

The sample size affects the sampling distribution too:

Standard error = population standard deviation / square root of sample size

$\sigma$Y-bar= $\sigma$ / $\sigma$ n

**Example for Standard error**

Standard error = population standard deviation / square root of sample size

$\sigma$ Y-bar= $\sigma$ / $\sigma$ n

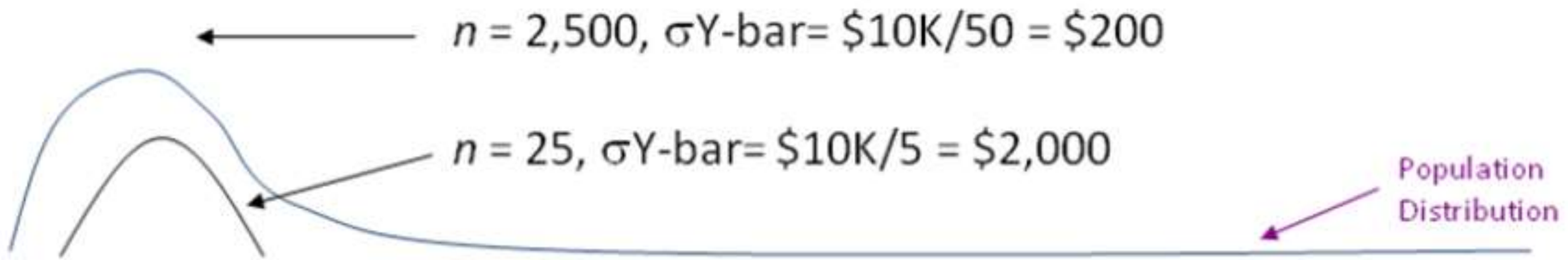
IF the population income were distributed with mean, $\mu$ = \$30K with standard deviation, $\sigma$ = \$10K

# What decides the Sampling Distribution?

Standard error = population standard deviation / square root of sample size

$$\sigma_{Y\text{-bar}} = \sigma/\sqrt{n}$$

IF the population income were distributed with mean, $\mu$ = \$30K with standard deviation, $\sigma$ = \$10K




...the sampling distribution changes for varying sample sizes

# So why are sampling distributions less variable when sample size is larger?

**Example 1:**

• Think about what kind of variability you would get

– *if you collected income through repeated samples of size 1 each.*

• Contrast that with the variability you would get:

– i*f you collected income through repeated samples of size N – 1 (or 300 million minus one) each.*

**Example 2:**

• Think about drawing the population distribution and playing “darts” where the mean is the bull’s-eye. Record each one of your attempts.

• Contrast that with playing “darts” but doing it in rounds of 30 and recording the average of each round.

• What kind of variability will you see in the first versus the second way of recording your scores.

**…Now, do you trust larger samples to be more accurate?**

# So why are sampling distributions less variable when sample size is larger?

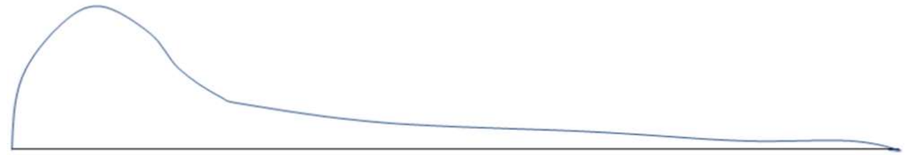
An Example:

A population's car values are $\mu$ = \$12K with $\sigma$ = \$4K.

Which sampling distribution is for sample size 625 and which is for 2500? What are their s.e.'s?

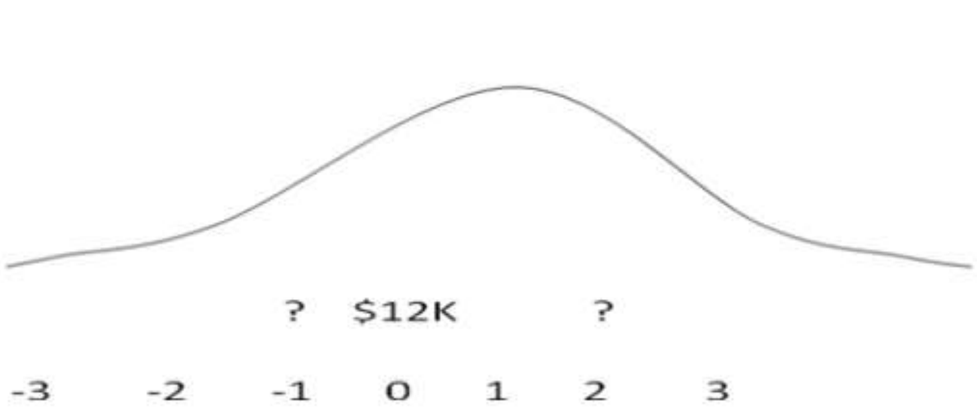
s.e. = \$4K/25 = \$160 s.e. = \$4K/50 = \$80

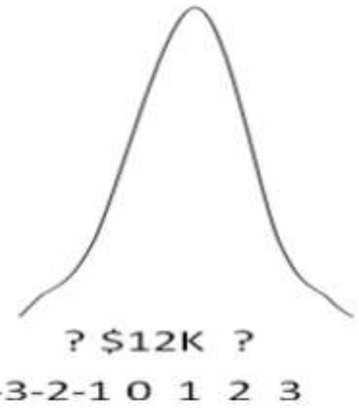
($\sqrt{625} = 25$) ($\sqrt{2500} = 50$)



A population's car values are $\mu$ = \$12K with $\sigma$ = \$4K.

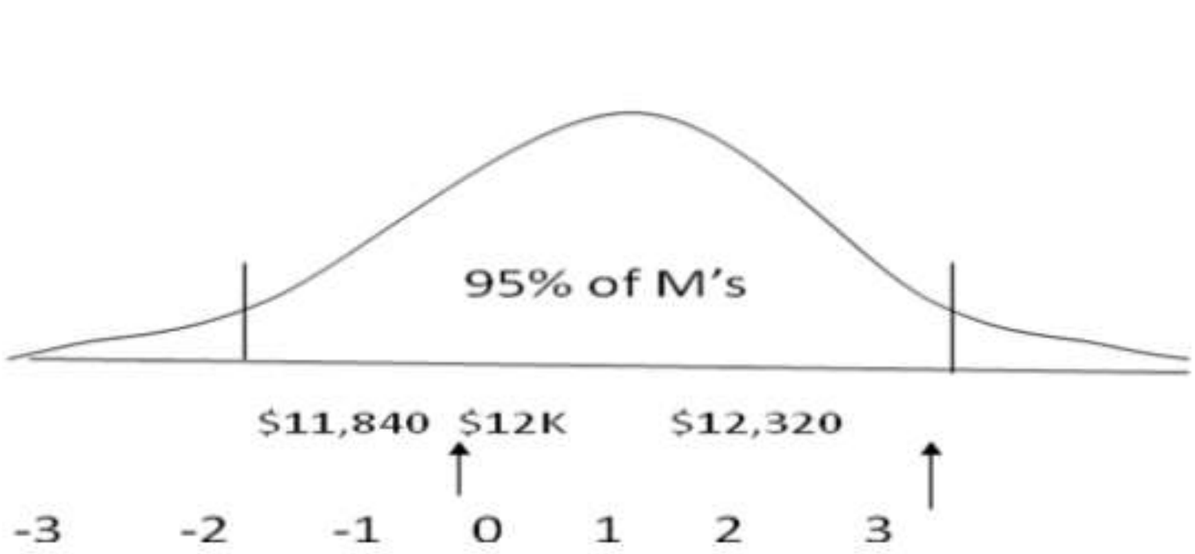
Which sampling distribution is for sample size 625 and which is for 2500? What are their s.e.'s?

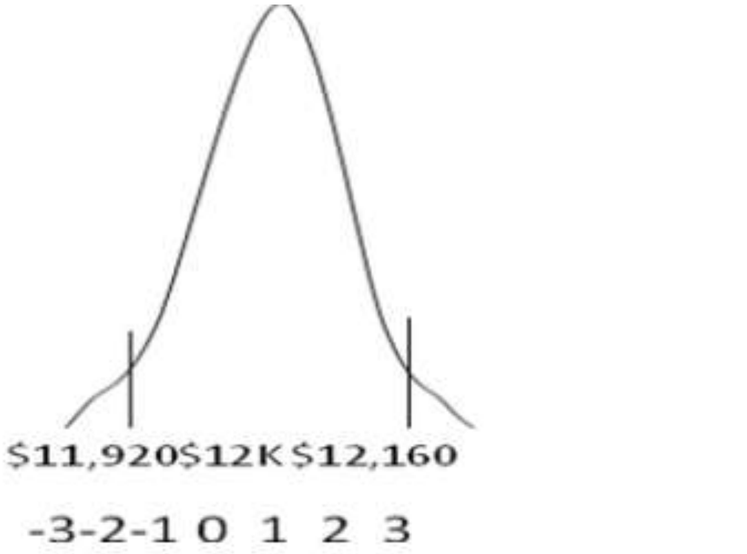


# So why are sampling distributions less variable when sample size is larger?

A population's car values are $\mu$ = \$12K with $\sigma$ = \$4K.
Which sampling distribution is for sample size 625 and which is for 2500?
Which sample will be more precise? If you get a particularly bad sample, which sample size will help you be sure that you are closer to the true mean?



# Some rules about the sampling distribution of mean

1. For a random sample of size *n from a population having mean* σ *and standard deviation* σ*, the sampling distribution of Y-bar (glitter-bar?) has mean* σ *and standard error* σ *Y-bar =* σ */* σ *n*
2. The Central Limit Theorem says that for random sampling, as the sample size *n grows, the sampling distribution of Y-bar approaches a normal distribution.*
3. The sampling distribution will be normal *no matter what the population distribution's shape as long as n > 30.*
4. If *n < 30, the sampling distribution is likely normal only if the underlying population's distribution is normal.*
5. As *n increases, the standard error (remember that this word means standard deviation of the sampling distribution) gets smaller.*
6. Precision provided by any given sample increases as sample size *n increases.*

# Some rules about the sampling distribution of mean

**Other Rules of Sampling Distribution**

So we know in advance of ever collecting a sample, that if sample size is sufficiently large:

- Repeated samples would pile up in a normal distribution
- The sample means will center on the true population mean
- The standard error will be a function of the population variability and sample size
- The larger the sample size, the more precise, or efficient, a particular sample is
- 95% of all sample means will fall between +/- 2 s.e. from the population mean

# Probability Distributions

- **A Note:**

– Not all theoretical probability distributions are Normal. One example of many is the binomial distribution.

- **The binomial distribution gives**

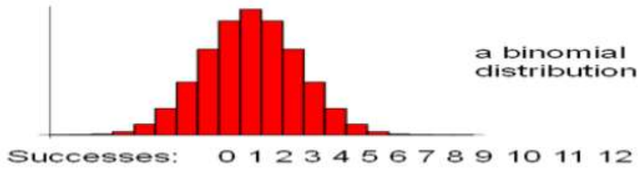
– the discrete probability distribution of obtaining exactly n successes out of N trials

- where **the result of each trial is true with known probability of success and false with the inverse probability.**
- The binomial distribution has

– a formula and

– changes shape with each probability of success and number of trials.

- However, in this class the normal probability distribution is the most useful!




• However, in this class the normal probability distribution is the most useful!

# POPULATION AND SAMPLING DISTRIBUTIONS

- Population Distribution
- Sampling Distribution

**Population Distribution**

Definition

The ***population distribution is the probability distribution of the population data.***

- Suppose there are only five students in an advanced statistics class and the midterm scores of these five students are:

70 78 80 80 95

- Let *x denote the score of a student*

Table 7.1 Population Frequency and Relative Frequency Distributions

# POPULATION AND SAMPLING DISTRIBUTIONS

Table 7.1 Population Frequency and Relative Frequency Distributions

| $x$ | $f$ | Relative Frequency |
|---|---|---|
| 70 | 1 | 1/5 = .20 |
| 78 | 1 | 1/5 = .20 |
| 80 | 2 | 2/5 = .40 |
| 95 | 1 | 1/5 = .20 |
| | $N = 5$ | Sum = 1.00 |

# POPULATION AND SAMPLING DISTRIBUTIONS

Table 7.1 Population Frequency and Relative Frequency Distributions

Table 7.2 Population Probability Distribution

| $x$ | P $(x)$ |
|---|---|
| 70 | .20 |
| 78 | 20 |
| | $\Sigma P(x) = 1.00$ |

Definition

The probability distribution of is called its sampling distribution.

It lists the various values that can assume and the probability of each value of . In general, the probability distribution of a sample statistic is called its ***sampling distribution.***

# References


www.statisticshowto.com/prediction-error-definition/

https://web.stanford.edu/class/msande226/lecture5_prediction.pdf

https://newonlinecourses.science.psu.edu/stat555/node/116/

https://online.stat.psu.edu/stat555/node/116/


thank
you!