



Module 4 : Sampling and Confidence Intervals



Sampling and Confidence Intervals

- ▶ Populations and Samples
- ▶ Point Estimation of a Population Mean and Proportion
- ▶ The Standard Normal
- ▶ Confidence Interval Estimation
- ▶ Sample Size Determination
- ▶ The Finite Correction Factor



Populations and Samples

- ▶ Definition of Population and Population Parameters

 - Populations and Samples

 - Population Parameters


- ▶ Samples and Sample Statistics

- ▶ Simple Random Sample

 - Sampling Strategies

 - Simple Random Sample

- ▶ Problems in Sampling (Bias)



Population vs Sample

- ▶ A population includes all elements
- ▶ CENSUS: survey the entire population => NO ERROR
- ▶ A sample: a subset of the population that observed
- ▶ More than one sample can be derived from the same population
- ▶ STATISTIC: A measurable characteristic of a sample
=> It's intended to represent or estimate the population
(with MINIMUM/ACCEPTABLE ERROR)

Population Parameters

- ▶ Population: collection of all objects of interest, example:
 - All voters registered for a US Presidential election
 - All the customers shopping in a department store on a given day
- ▶ Population Parameters: numerical characteristic of a population, example:
 - The fraction of voters that prefer the Democratic candidate for president.
 - The average weight of all the cows in India.
 - The standard deviation of the amount spent by a dept. store customer on a given day.



Sampling Strategy

- ▶ Sample must be representative of the population of interest.
- ▶ Stratification Method
- ▶ Random Method
- ▶ Cluster Method
- ▶ Multi-Stage Clustering Method
- ▶ Convenient Sampling Method
- ▶ Use random function in Excel to create a random number

Sampling Strategy – Simple Random Sample

Simple Random Sample (SRS)

- each set of n individuals has the same chance of being chosen
- Example Population $(N) = 5$ and sample size $(n) = 2$

Possible combination are:


(1,2) (1,3) (1,4) (1,5) (2,3) (2,4) (2,5) (3,4) (3,5) (4,5)

Each population member has chance $\mathbf{2/5 = 40\%}$ of being chosen

Sampling Strategy – Stratified Random Sampling

Stratified Random Sampling

- the population is divided into groups, called strata, based on some characteristic
- within each group, a sample, usually random, is selected
- How many are selected from each strata depends on the purpose for creating the strata initially
- In most cases, stratification is done to ensure that sample percentages match population percentages on some key characteristic



Sampling Strategy – Cluster Sampling

Cluster Sampling

- dividing up the population into clusters and then selecting clusters to be part of the sample
- Every cluster should represent the population on a small scale and be as heterogenous as possible
- Every population element must belong to one and only one cluster
- one-stage clustering (include all from the cluster)
- multi-stage clustering (randomly selects from the cluster)



Sampling Strategy – Systematic Random Sampling

Systematic Random Sampling

- Randomly picks the first item of the population
- Continue by picking the n^{th} subject from the list
- The results are usually representative of the population unless certain characteristics of the population are repeated for every n^{th} individual, which is highly unlikely



Sampling Strategy – Convenient Sampling

Convenient Sampling

- subjects are selected because of their convenient accessibility and proximity to the researcher
- Example Professor selected his students because they are easy to reach in campus

Problem in Sampling (Bias)

- ▶ Unintentional Error in preparing the sample:
 - Non Random Sample: 10th machine is faulty
 - Selection bias: survey by phone
 - Non-Response bias: reluctant to response
 - Publication bias: Bill Gates & Steve job => drop out from schooll
 - Response bias/ Desirability bias: response only what nice to say
 - Survivorship bias:

Point Estimation of a Population Mean and Proportion

- ▶ Mean, Variance and Standard Deviation of Sample Mean (\bar{x}):

 - Mean, Variance, and Standard Deviation of Sample Mean

 - Xbar and Mean, Variance, and Standard Deviation of Xbar

- ▶ Examples of Sample Mean (\bar{x})

 - Example of \bar{x} Formulas

- ▶ Estimate Population Proportion using P-hat (\hat{p})

Sample & Population Notations

σ^2 : Population variance

σ : Population standard deviation

s^2 : Sample variance

s : Sample standard deviation

μ : Population mean

\bar{x} : Sample mean

N : Number of observations in the population

n : Number of observations in the sample

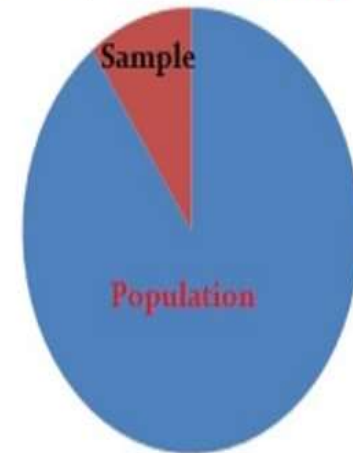
Population & Sample Notation

Notation	Population	Sample
Number of observation	N	n
Variance	σ^2	s^2
Standard Deviation	σ	s
Mean	μ	\bar{x}

Sample Mean

- ▶ The sample mean symbol is \bar{x} , pronounced “x bar”.
- ▶ The sample mean is an average value found in a sample.
- ▶ Formula: $\bar{x} = (\Sigma x_i) / n$ $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$
- ▶ The sampling distribution of the sample mean is a probability distribution of all the sample means.

The sample mean is an **average** value found in a sample.



Variance & Standard Deviation of Sample Mean

► Variance of Sample Mean:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Variance in a population is:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{n}$$

[x_i is the i th observation from a sample of the population, \bar{x} is the sample mean, n (sample size) - 1 is [degrees of freedom](#), Σ is the summation]

Standard deviation is Square root of variance

Standard Deviation of Sample

- ▶ Standard Deviation of Sample Mean:

$$(\hat{p}) = \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

\hat{p} : point estimation of p

\bar{x} : point estimation of μ



The Standard Normal & the .S Functions

- ▶ The standard normal distribution is special case of the normal distribution.
- ▶ It is a distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.
- ▶ Often called as Z
- ▶ NORM.S Functions: Excel function take into account the $\mu=0$ and $\sigma=1$



Confidence Interval Estimation

- ▶ Confidence Interval for Population Mean:
 - to describe the amount of uncertainty associated with a sample estimate of a population parameter
 - Examples: 95% Confidence Interval for Population Mean
- ▶ Demonstration of Meaning for Confidence Interval:
 - Demonstration of Meaning for 95% Confidence Interval
- ▶ Confidence Interval for Population Proportion
 - 95% Confidence Interval for Population Proportion
 - Blyth's Formula for Proportion Confidence Interval

Confidence Interval Estimation

► Demonstration of Meaning for 95% Confidence Interval:

You are told the standard deviation of invoice values is \$500.

A sample of 100 invoices taken from a large sample of invoices has a sample mean value of \$4500.

You are 95% sure the mean size of an invoice is within what range?

	C	D	E
1			
2		samplemean	4500
3		popsigma	500
4		samplesize	100
5		z.025	-1.95996
6		z.975	1.959964
7			
8		Lower Limit	4402.002
9		Upper Limit	4597.998



Sample Size Determination

- ▶ The bigger sample size, the better quality of Population mean estimation
- ▶ Sample Size for Estimating Population Mean
- ▶ Sample Size for Estimating a Population Proportion

Sample Size for Estimating Population Mean

► M4L5HW1

In a Presidential election poll, how many voters need to be sampled so we can be 95% confident that our estimate is within 2% of the true percentage of voters preferring the Democratic candidate?.

► Solution/Answer:

$$n = 1.96^2 / (4 * E^2)$$

$$n = 2400.912 \approx 2401$$

(Must be rounded UP)

	A	B	C	D
1	<u>M4L5HW1</u>	$n = 1.96^2 / 4E^2$		
2	Error	0.02		
3	α	0.05		
4	$Z_{\alpha/2}$	1.959963985		
5	Sample Size	2400.911763	$= Z_{\alpha/2}^2 / (4 * Error^2)$	
6		≈ 2401		

Sample Size for Estimating Population Mean

- ▶ If the standard deviation (σ) is KNOWN

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$$

The formula above provide the sample size needed (n) under the requirement of population mean interval estimate at $(1 - \alpha)$ confidence level, margin of error E , and population variance σ^2 .

$z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

Sample Size for Estimating Population Mean

► If the standard deviation (σ) is UNKNOWN

a. $\sigma \approx \text{range}/4$

Why $1/4$? (the highest value of $(p * (1-p))$ happens when $p=0.5$ is $1/4$)

$$(1-p) * p = 0.5 * 0.5 = 1/4$$

$$n = (Z_{\alpha/2})^2 / (4 * E^2)$$

–OR–

a. Calculate sample standard deviation (s) and use it in place of σ

b. Estimate the value of σ using other method

Sample Size for Estimating Population Mean

► Problem

Assume the population standard deviation σ of the student height in survey is 9.48. Find the sample size needed to achieve a 1.2 centimeters margin of error at 95% confidence level.

► Solution/Answer:

$$n = 239.75 \approx 240$$

	A	B	C	D
1	Error	1.2		
2	α	0.05		
3	$Z_{\alpha/2}$	1.959964	=NORM.S.INV(1-Alpha/2)	
4	sigma	9.48		
5	Sample Size	239.7454	=Zalpha2^2*sigma^2/Err^2	
6		≈ 240		

Sample Size for Estimating Population Mean

► **M4L5HW2:**

How many soda cans need to be sampled in order to be 95% confident that your estimate of the average number of ounces in a soda can is accurate within 0.03 ounces?

Assume standard deviation of ounces in a can is 0.15 ounces.

► **M4L5HW3:**

How many American men's heights need to be sampled to be 95% confident that you can estimate the average height of an American man within 1 inch?

Assume the standard deviation of an American men's height is 3 inches.

Sample Size for Estimating Proportion

► Sample

suppose we want to estimate the fraction of registered voters preferring the Republican candidate in a Texas election. We would like our estimate to have a 95% chance of being accurate within 3%.

How large of a sample is needed?

Answer: 1067.072

$$n = 1.96^2 / (4 * E^2)$$

$$n = 1068$$

(Must be rounded UP)

	A	B	C	D
1	Error	0.03		
2	α	0.05		
3	$Z_{\alpha/2}$	1.959964	=NORM.S.INV(1-Alpha/2)	
4	sigma	0.5		
5	Sample Size	1067.072	=Zalpha2^2*sigma^2/Err^2	
6		≈ 1068		

The Finite Correction Factor

- ▶ Finite Correction Formula for Sample Size

Applied when sample size bigger (> 10% of population)

n: sample size with Correction factor

N_0 : sample size without Correction factor

N: Population size

$$n = \frac{N_0 * N}{N_0 + N - 1}$$



The Finite Correction Factor

► Sample:

Suppose we want to estimate the mean salary of Fortune 500 CEOs and be 95% sure our estimate is accurate within \$1 million. How large of a sample is needed? (standard deviation is known \$5 million)

The Sample Size with Finite Correction Factor

► Solution:

	A	B	C	D	E	F
1	Error	1				
2	α	0.05				
3	$Z_{\alpha/2}$	1.959964	=NORM.S.INV(1-Alpha/2)			
4	sigma	5				
5	N	500				
6	Sample Size(No FC)	96.03647	=Zalpha2^2*sigma^2/Err^2			
7	Sample Size(With FC)	80.69797	=SampleNoFC*PopSize/(SampleNoFC+PopSize-1)			
8		≈ 81				

The Finite Correction Factor

► Finite Correction Formula for Estimating Population Mean

Applied when sample size bigger (> 10% of population)

n: sample size with Correction factor

N_0 : sample size without Correction factor

N: Population size

$$n = \frac{N_0 * N}{N_0 + N - 1}$$

Finite Correction Factor Confidence Interval for Population Mean

► Sample:

Suppose we want to estimate the average salary of Fortune 500 CEOs.

Assume the standard deviation of these salaries is known to be \$5 million.

If we sample 100 CEOs and find an average salary of \$40 million,

With 95% confidence interval, we are sure that the actual mean salary of Fortune 500 CEOs is between \$_____ and \$_____

Finite Correction Factor Confidence Interval for Population Mean

► Solution: between \$39.12 to \$40.87

	A	B	C	D	E	F
1						
2						
3	samplesize (n)	100				
4	popsiz(N)	500				
5	sigma	5				
6	xbar	40				
7						
8	FC	0.895323				
9	With FC					
10	LowerLimit	39.12258	$=xbar-1.96*FC*sigma/SQRT(samplesize)$			
11	UpperLimit	40.87742	$=xbar+1.96*FC*sigma/SQRT(samplesize)$			
12						
13	Without FC					
14	LowerLimit	39.02	$=xbar-1.96*sigma/SQRT(samplesize)$			
15	UpperLimit	40.98	$=xbar+1.96*sigma/SQRT(samplesize)$			



The Blyth Formula

- ▶ Blyth formula: a confidence interval formula to be used when all trials result in success or failure (one outcome is nearly 100%, the other one is nearly 0%)
- ▶ Suppose my son has driver to work 500 times without an accident. Let's find a 95% confidence interval for the chance he will have, or will not have an accident.

The Blyth Formula

Simply enter the number of trials in cell C3 and alpha of .05 for a 95% confidence interval (alpha of .01 for a 99% confidence interval) in C4. In cells C8 and D8 we find that we are 95% sure the chance of an accident is between 0 and 0.005974 and from cells C11 and D11 we find that we are 95% sure the chance of no accident is between 0.99492645 and 1.

	A	B	C	D	E	F
1	Blyth Confidence Interval					
2						
3		n	500			
4		alpha	0.05			
5						
6						
7		Successes	Lower	Upper		
8		0	0	0.005974		$=1-\alpha^{(1/n)}$
9		1	5.0634E-05	0.007351	$=1-(1-0.5*\alpha)^{(1/n)}$	$=1-(0.5*\alpha)^{(1/n)}$
10		499	0.99264939	0.999949	$=(0.5*\alpha)^{(1/n)}$	$=(1-0.5*\alpha)^{(1/n)}$
11		500	0.99402645	1	$=(\alpha)^{(1/n)}$	1

The Finite Correction Factor

► **M4L6HW1:**

In a local election poll in a town with 5000 registered voters, how many voters need to be sampled so we can be 95% confident that we can estimate the true percentage of voters preferring the Democratic candidate within 2%? Apply the finite correction factor to obtain your answer..

► **Answer:**

The Finite Correction Factor

► **M4L6HW2:**

In a batch of 200 soda cans, how many soda cans need to be sampled in order to be 95% confident that your estimate of the average number of ounces in a soda can is accurate within 0.03 ounces? Assume standard deviation of ounces in a can is 0.15 ounce.

Apply the finite correction factor to obtain your answer.

► **Answer:**

The Finite Correction Factor

► **M4L6HW3:**

You are told the standard deviation of the invoice size in a population of 200 invoice values is \$1,000. A sample of 50 invoices yields an average invoice size of \$5,000. You are 95% confident that the average size of an invoice is between which two values? Calculate the lower and upper limit, and apply the finite correction factor to obtain your answer..

► **Answer:**

Homework & Quiz