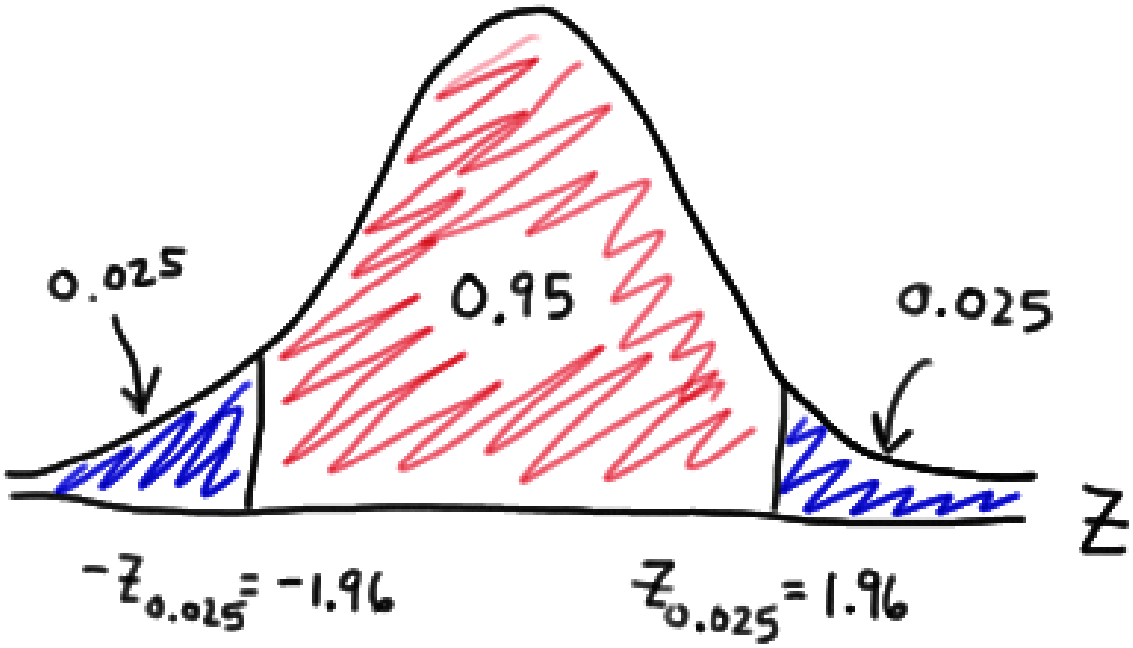
SAMPLING AND CONFIDENCE INTERVALS

DR. ALVIN ANG



CONTENTS

PART I

SAMPLING

A. WHY IS SAMPLING NECESSARY

1. Too Time Consuming
    1. To get statistics for whole population takes too much time.

2. Too Expensive
    1. Exorbitant to pay surveyors for entire population.

3. Impossible to get statistics for whole population.

4. Tests may be Destructive.
    1. The manufacturer of fuses cannot test all of them because in the testing the fuse is destroyed and none would be available for sale.

5. Not Necessary to Get Statistics of Entire Population.
    1. If the samples statistics show credible results, sampling more probably would not change the results significantly.

B. RESULTS OF OVER / UNDER-SAMPLING

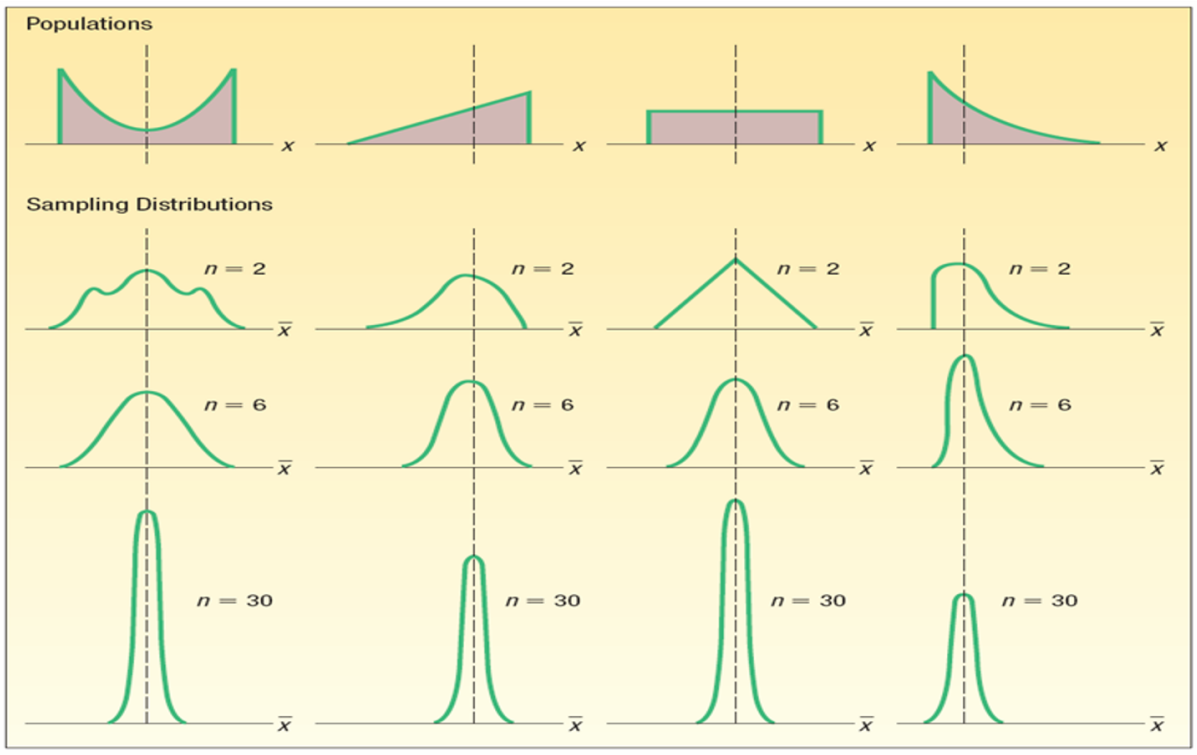


CHART 8–2 Results of the Central Limit Theorem for Several Populations

- Figure above shows the Central Limit Theorem (CLT) giving rise to the Normal Distribution (Ang, 2020).
- At the top, we have the true population distributions.
- According to the CLT, as we progress from a sample size of 2 to 30, we see that the Normal Distribution curve begins to form nicely around the median of the Population Distribution.
- However, should we progress from n = 30 to 300 to 3000 to 3,000,000 (or rather, we have "sampled" the entire "population")…. The curve goes back to looking like the Population Distribution.
- In other words, we cannot under nor over sample.
- We need an appropriate sample size.

C. TYPES OF SAMPLING: PROBABILITY VS NON-PROBABILITY SAMPLING

1. PROBABILITY SAMPLING

Simple Random Sampling

Definition

- Just anyhow… put paper slips in box and shuffle…
- Population size N
- Sample size n
- Every member has same probability of being selected = n /N

Advantages

- Easy to generate and use
- Easy to understand

Disadvantages

- Rarely used in real applications
- May result in under- or over-representation of certain segments of the population.

Systematic Random Sampling

Definition

- Numbering System
- Population size N
- Sample size n
- Systematic sampling = N/n
- It randomly selects one member from the first group and then takes every N/n member.

Example

- N = 55, 000 names in a telephone directory
- n = sample of 250 names.
- The sample interval is 55, 000 / 250 = 220.
- Choose one name randomly from name 1 to name 220.
- Then take every 220th name in the book.
- E.g. 3, 223, 443, 663…

Advantages

- Cheap
- Easy to implement

Disadvantages

- If the interval was 7, then we would end up selecting the same day each week for our sample.

Stratified Random Sampling

Definition

- Classify based on characteristics (or Strata)

Example

- Stratify all Army personnel (Population = 600 pax) into
    - Strata 1: Total Generals = 180
    - Strata 2: Total Officers = 120
    - Strata 3: Total Enlisted Personnel = 300
- Sample Size = 50 pax
- Proportioning
    - Generals = 30% (=0.3 x 50 = 15 Generals)
    - Officers = 20% (=0.2 x 50 = 10 Officers)
    - Enlisted Personnel = 50% (=0.5 x 50 = 25 Enlisted)

Advantages

- Easy to implement
- Can obtain information about each strata separately
- Less variability within a strata than the population

Disadvantages

- Ignore the possibility of some strata having high variability (which means we should sample more from these to take account of increased variability.

Cluster Sampling

Definition

- Similar to Stratified
- Classified based on Geographical Boundaries.

Example

- Geographic Boundaries
- Clusters of Schools

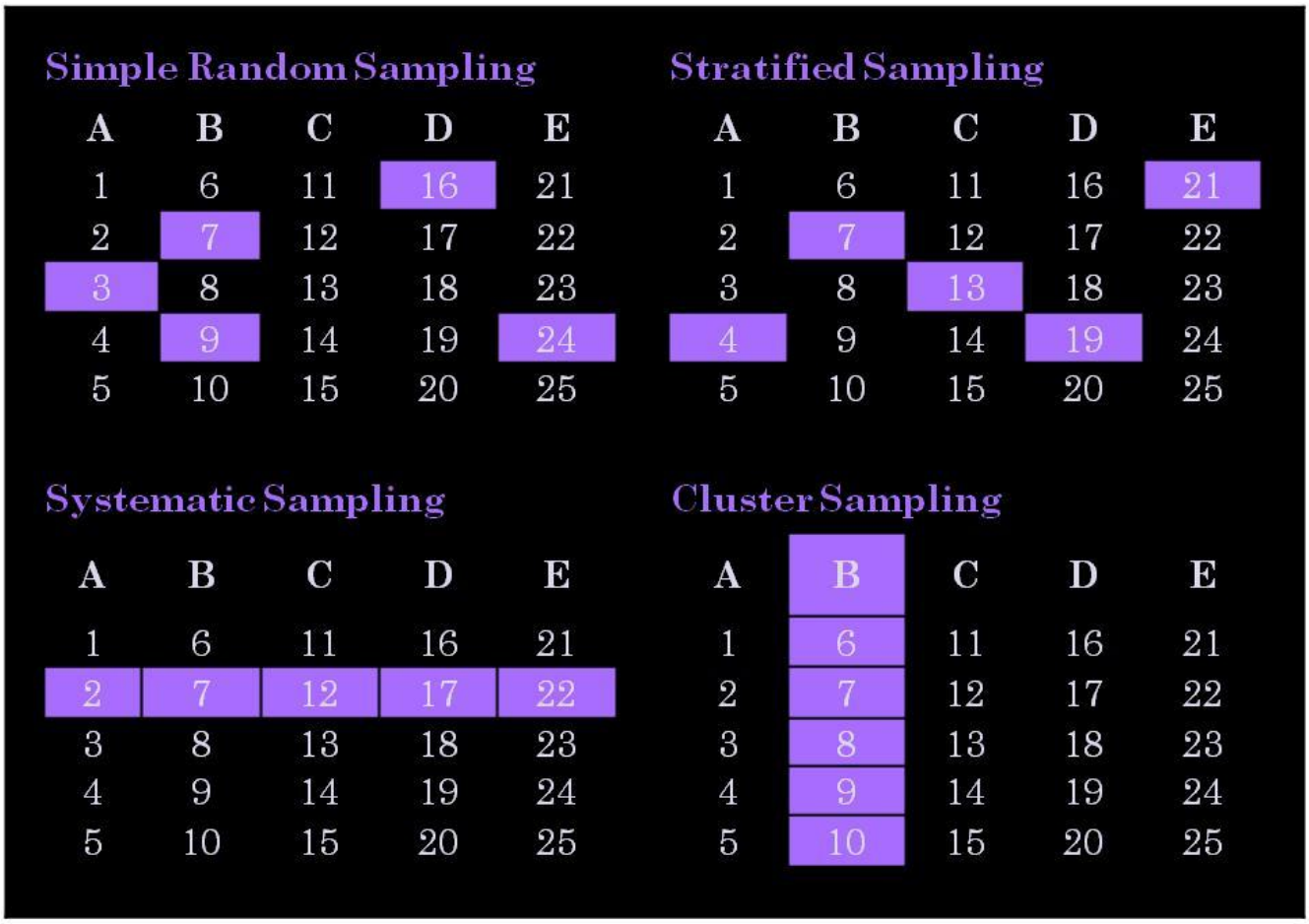
Advantages

- Provides an unbiased estimate of population parameters if properly done
- Economically more efficient than simple random
- Lowest cost per sample

Disadvantages

- Often lower statistical efficiency due to subgroups being homogeneous
- Moderate cost

Summary of Probability Sampling Methods



2. NON-PROBABILITY SAMPLING – PURPOSIVE VS NON-PURPOSIVE SAMPLING

Purposive Sampling

Judgement Sampling

- Arbitrarily selects sample units to conform to some criterion.
- This is appropriate for the early stages of an exploratory study.

Quota Sampling

- Sample based on relevant characteristics that describe some dimensions of the population.
- Researchers may specify more than one control dimension.
- Each dimension should have a distribution in the population that is pertinent to the topic studied.

Non-Purposive Sampling

Convenience Sampling

- Based on ease of accessibility.
- Least reliable but cheapest
- Easiest to conduct.
- Examples include informal pools of friends and neighbors, people responding to an advertised invitation, and "on the street" interviews.

Snowball Sampling

- Participants are referred by the current sample. (Referral networks).
- Used frequently in qualitative studies.

3. PROS AND CONS OF PROBABILITY VS NON-PROBABILITY SAMPLING

	Advantages	Disadvantages
Non-probability Sampling	☑ Least expensive and time consuming ☑ Sampling units are accessible, easy to measure and co-operative ☑ Random aspect may be able to be used	☒ Selection bias ☒ May not be representative of population ☒ Cannot generalise to the population
Probability Sampling	☑ Easily understood ☑ Results may be projected to target population ☑ Different techniques (e.g. stratified) can help improve SRS results ☑ Stats tools applicable	☒ May be time-consuming, difficult and costly to do ☒ May be difficult to construct sample frame that will permit a SRS ☒ SRS can result in large samples ☒ SRS may not result in as precise or representative sample ➔ may need to use a more complex technique (e.g. stratified, cluster, two-stage)

D. SAMPLING ERROR

TYPES OF ERROR

Sampling Error

- We took an "unlucky" sample
- E.g. extreme results
- Sample mean will be too large (or too small) an estimate for the population mean
- We can measure the Sampling Error using = $\overline{X} - \mu$ &
- Standard Error of the Sample Mean: $s_{\overline{X}} = \frac{s}{\sqrt{n}}$
- More about this in the next section.

Non – Sampling Error

- We cannot measure non-sampling error.
- We can only control it through good sampling techniques and survey instruments.
- Examples of Non-Sampling Error:
    - No response: Members who fail to provide requested information
    - Non truthful responses: Members lie, particularly about sensitive information
    - Measurement error: poorly worded questions lead to poor responses

First, defining the Sampling Error…

Since:

- μ: Population Mean (usually unknown)
- σ: Population Std. Dev. (usually unknown)
- $\overline{X}$: Sample Mean (best estimate of μ)
- s: Sample Std. Dev.
- Sampling Error = $\overline{X} - \mu$
- Standard Error of the Population Mean: $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$
- Since we usually don't have the Population Std. Dev. σ, Standard Error of the Sample Mean: $s_{\overline{X}} = \frac{s}{\sqrt{n}}$
- So actually, the {Sampling Error} is = the {Standard Error of the Sample Mean}!
- This is because, as we go around taking different samples, the {Sampling Error} changes. But the average of all these changes will = $s_{\overline{X}} = \frac{s}{\sqrt{n}}$

Next, defining the Confidence Interval (CI)…

Confidence Interval for the Population Mean, σ unknown	$\bar{X} \pm t\frac{s}{\sqrt{n}}$

Where $\bar{X}$ is the sample mean.
t is the value associated with the given level of confidence.
s is the sample standard deviation.
n is the size of the sample.

Figure 1: CI for Population Mean, Sigma unknown (SUSS, 2014)

- Although we will only discuss about CI in the next chapter, notice that the above equation is actually
- $\bar{X} \pm t \times Std.\ Err.\ of\ \ Sample\ Mean$
  $= \bar{X} \pm ts_{\bar{X}}$
- This is how the {Standard Error of the Sample Mean} gave birth to the CI.

Example of how Sampling Error gives birth to Confidence Interval (CI)…

Given:

- Population size = 10,000
- Sample size = 100
- $\overline{X} = \$285$
- s = $43
- $s_{\overline{X}}$ = $4.30

Find:

- CI

Since:

- Referring to t distribution table, 95% CI
- Degree of Freedom = n – 1 = 99
- t = 1.984 ≈2

Thus:

- The CI is actually $\overline{X} \pm ts_{\overline{X}} = \$285 \pm (2)(\$4.30)$
  $= \$285 \pm \$8.60$
- In layman terms, we are 95% confident that the Population Mean is between $276.40 and $293.60.

E. HOW TO CALCULATE SAMPLE SIZE?

SAMPLE SIZE FOR ESTIMATING POPULATION MEAN

Sample Size for Estimating the Population Mean	$n = \left(\frac{z\sigma}{E} \right)^2$

Figure 2: Sample Size for Estimating the Population Mean (SUSS, 2014)

where

n is the size of the sample.

z is the standard normal value corresponding to the desired level of confidence.

σ is the population standard deviation.

E is the maximum allowable error.

Points to Note:

- If σ is large → it means that large variability → then we need to increase n (sample size) to lower it.
- But if we don't know σ, how to calculate n then?
- Answer: Estimate σ.
- How to estimate?
    - Use comparison: find a comparable study where its σ is available.
    - Use $\sigma \approx \frac{R}{6}$; where R is the Range = Max – Min.
    - Use $s \approx \sigma$; meaning to find a small sample and calculate its s, sample std. dev. And approximate it to σ.

Example:

Given:

- σ ≈ 3 minutes
- Allowable Error, E = ±1 minute

Find:

- Required Sample Size, n
- For finding the mean time of customer spending in store
- With 95% CI
- Means Z = 1.96

Level of Significance (Alpha)	Critical value (Z_α) of Z	
	Two-tailed test	Single tailed test
10%	1.645	1.28
5%	1.96	1.645
1%	2.58	2.33

Answer:

- $n = \left(\frac{z\sigma}{E} \right)^2$
- $n = \left(\frac{(1.96)(3)}{1} \right)^2 = 35$

SAMPLES SIZE FOR ESTIMATING POPULATION PROPORTION

What is the Sample Proportion?

Sample Proportion	$p = \frac{X}{n}$

Where p is the sample proportion

X is the number of successes in the sample.

n is the number of items sampled.

**Note: Sometimes, p is not given. When no estimate of p is given, we assume it to be 0.50.

What is the Sample Size Required for Population Proportion?

Sample Size for the Population Proportion	$n = p(1-p)\left(\frac{z}{E}\right)^2$

Where:

P is the estimated proportion based on the pilot survey.

z is the z score associated with the degree of confidence selected.

E is the allowable error.

Example:

Given:

- p : Population Proportion = 0.15 (15% of the population was unemployed)
- E: Allowable Error = 0.05
- 95% Confidence → Z = 1.96

Find:

- The government wants to be 95% confident that 15% of the population was unemployed.
- What should the sample size, n, be?

Answer:

- $$n = p(1-p)\left(\frac{z}{E}\right)^2 = 0.15(1-0.15)\left(\frac{1.96}{0.05}\right)^2 = 195.92 = 196$$

PART II

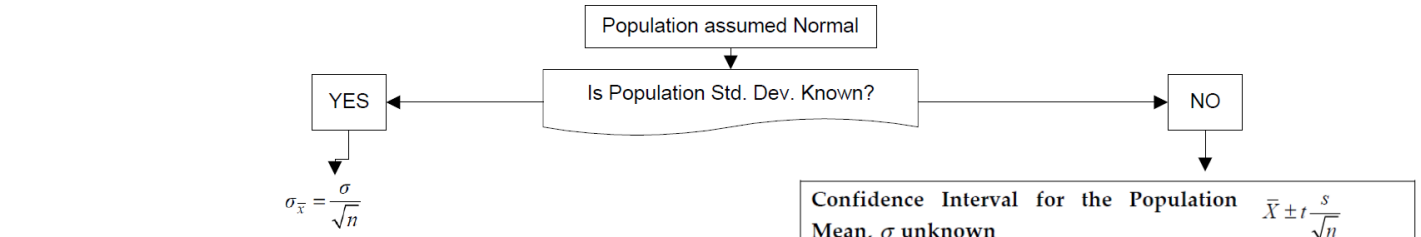
CONFIDENCE INTERVALS (CI)

A. DEFINITION

- A confidence interval is a range of values.
- It is constructed from the sample data.
- The population parameter is likely to occur within that range at a specified probability.
- This specified probability is called level of confidence.

The level of confidence is a measure of the confidence we have that an interval estimate will include the population parameter.

B. WHEN TO USE WHICH



$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Where $\sigma_{\bar{x}}$ is the symbol for the standard error of the mean.
σ is the population standard deviation.
n is the number of observations in the sample.

Confidence Interval for the Population Mean with σ Known $\overline{X} \pm z\frac{\sigma}{\sqrt{n}}$

Where $\overline{X}$ is the sample mean.
z is the critical value from the distribution (in this case, the Normal) and depends on the level of confidence.
σ is the population standard deviation.
n is the size of the sample.

Confidence Interval for the Population Mean, σ unknown $\bar{X} \pm t\frac{s}{\sqrt{n}}$

Where $\bar{X}$ is the sample mean.
t is the value associated with the given level of confidence.
s is the sample standard deviation.
n is the size of the sample.

C. CI FOR Z

EXAMPLE:

Given:

- Sample size, n = 64 trucks
- Sample mean, $\overline{X} = \$1,200$
- Population Standard Deviation, $\sigma = \$280$
- Assume Normal Distribution

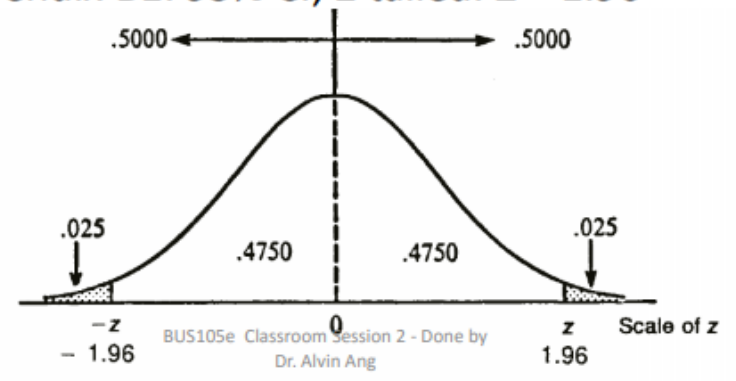
Find:

- 95% Confidence Interval

Answer:

Confidence Interval for the Population Mean with σ Known $\overline{X} \pm z\frac{\sigma}{\sqrt{n}}$

- See Appendix B1: 95% CI; 2 tailed: Z = 1.96



$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}} = \$1,200 \pm 1.96\frac{\$280}{\sqrt{64}}$$
$$= \$1,200 \pm \$68.6$$
$$= \$1,131.40 \quad to \quad \$1,268.60$$

- This means:
- *"We are 95% Confident that Average Repair Cost will be between $1131 and $1268"*

D. CI FOR T

EXAMPLE:

Given:

- Sample size, n = 12 batteries
- Sample mean, $\overline{X} = 4.32 hours$
- Sample Standard Deviation, $s = 0.1469 hours$
- Assume Normal Distribution

Find:

- 95% Confidence Interval
- Population Std. Dev. Is unknown

Answer:

Confidence Interval for the Population Mean, σ unknown $\overline{X} \pm t\frac{s}{\sqrt{n}}$

- Refer Appendix B2: 95% CI; Degree of Freedom (df) = 12 – 1 = 11 → t = 2.201

$$\overline{X} \pm t\frac{s}{\sqrt{n}} = 4.32 \pm 2.201\frac{0.1469}{\sqrt{12}} = 4.32 \pm 0.09332$$

Extra Question:

Is it reasonable to claim X bar = 4.25 hrs?

I.e. is it possible for battery to last 4.25 hrs?

Answer:

- 4.32 – 0.09332 = 4.2269
- 4.32 + 0.09332 = 4.4133
- Since 4.2269 < 4.25 < 4.4133
- We are 95% Confident battery can last 4.25 hours

BUS105e Classroom Session 2 - Done by Dr. Alvin Ang

Extra Question:

Is it reasonable X bar = 5 hours?

I.e. Battery last 5 hours?

Answer:

- Since 5 > 4.4133
- We are 95% Confident battery ***cannot*** last 5 hours

Compute Confidence Interval

Enter the data in Excel.

Click **Data → Data Analysis → Descriptive Statistics → OK**.

The following dialog box will appear. Follow the steps indicated below.

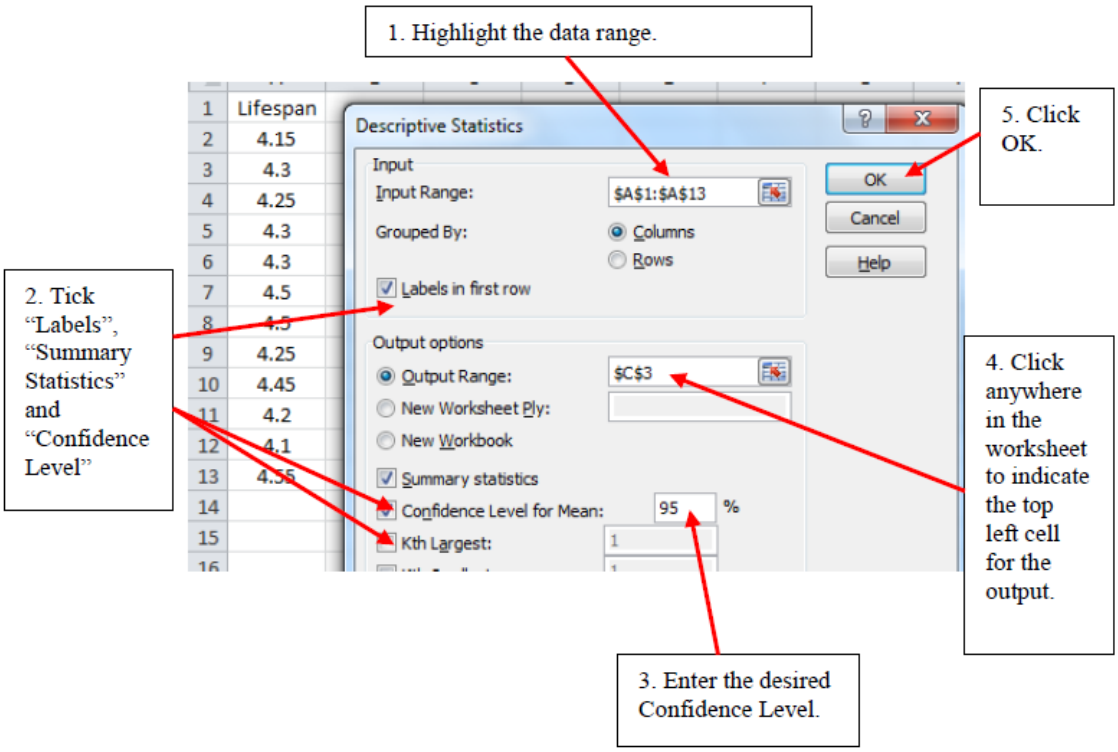
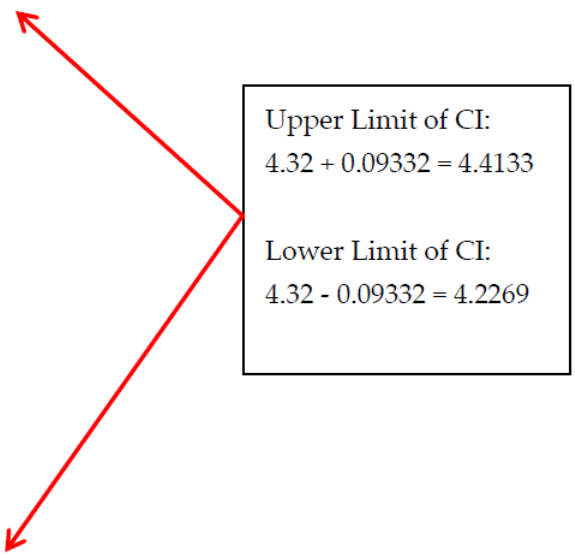


Figure 3: Using Excel to obtain the CI (SUSS, 2014)

Lifespan	
Mean	4.320833
Standard Error	0.042399
Median	4.3
Mode	4.3
Standard Deviation	0.146874
Sample Variance	0.021572
Kurtosis	-1.12413
Skewness	0.256023
Range	0.45
Minimum	4.1
Maximum	4.55
Sum	51.85
Count	12
Confidence Level(95.0%)	0.093319



Therefore, the 95% confidence interval for the population mean is from 4.2269 to 4.4133.

Figure 4: 95% CI using Excel (SUSS, 2014)

F. CI FOR POULATION PROPORTION

Sample Proportion $p = \frac{X}{n}$

Where p is the sample proportion
X is the number of successes in the sample.
n is the number of items sampled.

Standard Error of the Sample Proportion $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$

Confidence Interval for a Population Proportion $p \pm z\sigma_p$

Where: p is the sample proportion
σ_p is the "standard error" of the proportion.

Figure 5: Formulas for population proportion (SUSS, 2014)

EXAMPLE:

Given:

- X: Number of homeowners willing to switch electrical supplier = 800
- n: sample size = 1200 homeowners
- Confidence Level = 99%

Find:

- The Population Proportion, p
- The Standard Error of the proportion
- 99% CI for Population Proportion

- Interpret the results if the legislator state that 2.3 of the homeowners would switch.

Answer:

- the population proportion. $p = \frac{X}{n} = \frac{800}{1200} = 0.667$
- The standard error of the proportion is:
$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.667(1-0.667)}{1200}} = 0.0136$$
- The 99 percent confidence interval is found by:
$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$
$$= 0.667 \pm 2.58\sqrt{\frac{0.667(1-0.667)}{1200}}$$
$$= 0.667 \pm 2.58 \times 0.0136$$
$$= 0.667 \pm 0.035$$
$$= 0.632 \text{ and } 0.702$$

- The value 0.667 is in the interval, thus the legislator is correct in stating that 2/3 of the homeowners would switch energy suppliers.

REFERENCES

Ang, D. A. (2020). Probability Models.

SUSS. (2014). *BUS105e Study Guide - Business Statistics*. Singapore: Singapore University of Social Sciences (SUSS).

ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.