SAMPLING METHODS

UTKARSH KULSHRESTHA

kuls.utkarsh1205@gmail.com

Retail Kloud9 Technologies

NLP Engineer

Populations vs. Samples

- Who = Population:
  - all individuals of interest
  - US Voters, Dentists, College students, Children
- What = Parameter
  - Characteristic of population
- Problem: can't study/survey whole pop
- Solution: Use a sample for the "who"
  - subset, selected from population
  - calculate a statistic for the "what"

Types of Sampling

- Simple Random Sampling
- Stratified Random Sampling
- Cluster Sampling
- Systematic Sampling
- Representative Sampling
  (Can be stratified random or quota sampling)
- Convenience or Haphazard Sampling
- Sampling with Replacement vs. Sampling without Replacement

Types of Statistics

- Descriptive statistics:
  - Organize and summarize scores *from samples*

- Inferential statistics:
  - Infer information *about the population* based on what we know from sample data
  - Decide if an experimental manipulation has had an effect

THE BIG PICTURE OF STATISTICS

Theory

Question to answer / Hypothesis to test

Design Research Study

Collect Data
(measurements, observations)

Organize and make sense of the #s

USING STATISTICS!

Depends on our goal:

Describe characteristics
organize, summarize, condense data

Test hypothesis, Make conclusions, *interpret* data, understand relations

DESCRIPTIVE STATISTICS

INFERENTIAL STATISTICS

Types of Variables

- Quantitative
  - Measured in amounts
  - Ht, Wt, Test score
- Qualitative
  - Measured in categories
  - Gender, race, diagnosis

- Discrete:
  - separate categories
  - Letter grade
- Continuous:
  - infinite values in between
  - GPA

Scales of Measurement

- Nominal Scale: Categories, labels, data carry no numerical value
- Ordinal Scale: Rank ordered data, but no information about distance between ranks
- Interval Scale: Degree of distance between scores can be assessed with standard sized intervals
- Ratio Scale: Same as interval scale with an absolute zero point.

SAMPLING

- A sample is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” (Field, 2005)
- 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

SAMPLING……

- What is your population of interest?
  - To whom do you want to generalize your results?
    - All doctors
    - School children
    - Indians
    - Women aged 15-45 years
    - Other
- Can you sample the entire population?

SAMPLING…….

- 3 factors that influence sample representative-ness
  - Sampling procedure
  - Sample size
  - Participation (response)

- When might you sample the entire population?
  - When your population is very small
  - When you have extensive resources
  - When you don’t expect a very high response

Who do you want to generalize to?

The Theoretical Population

What population can you get access to?

The Study Population

How can you get access to them?

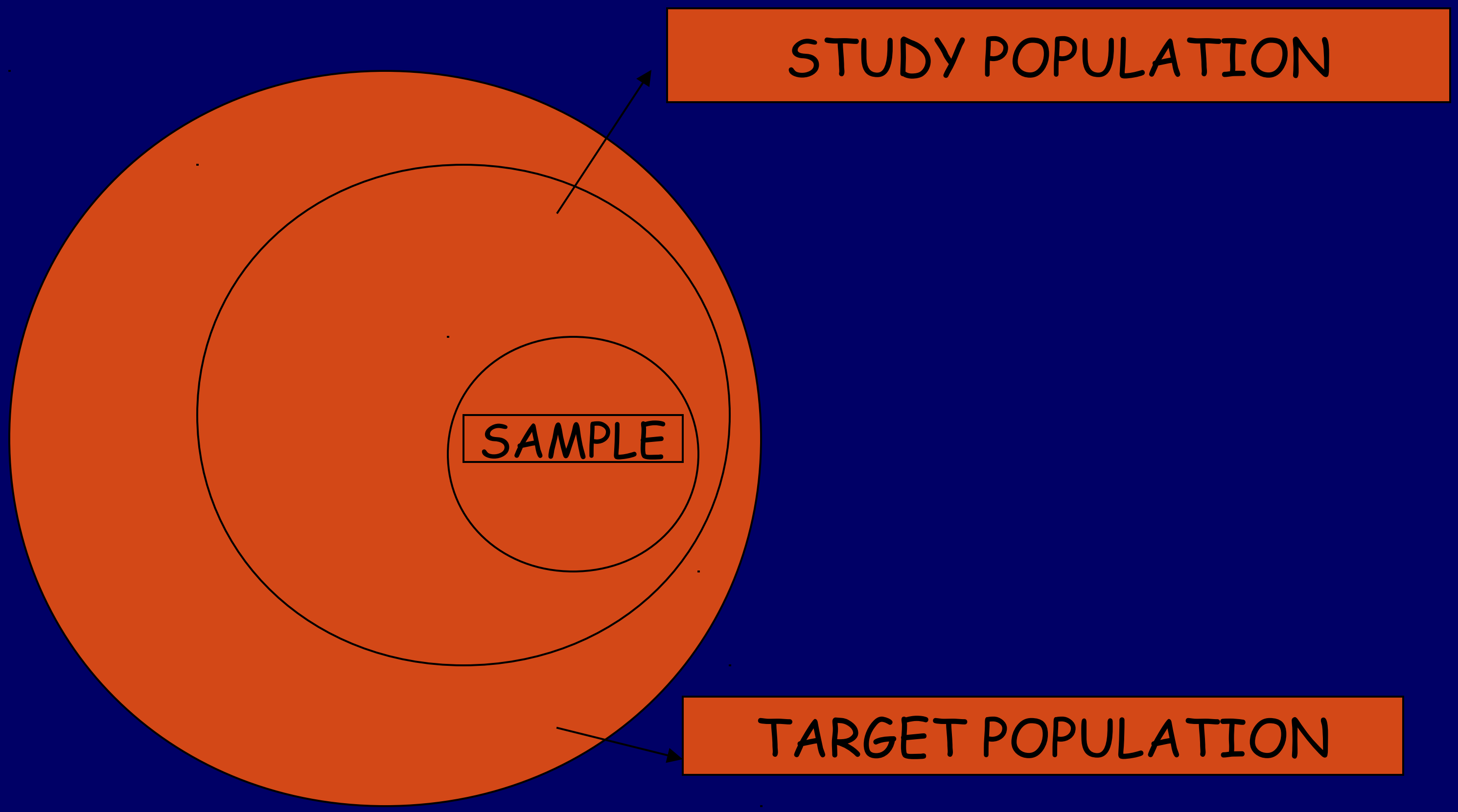
The Sampling Frame

Who is in your study?

The Sample

SAMPLING BREAKDOWN

SAMPLING.......



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

SIMPLE RANDOM SAMPLING

- Applicable when population is small, homogeneous & readily available
- All subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection.
- It provides for greatest number of possible samples. This is done by assigning a number to each unit in the sampling frame.
- A table of random number or lottery system is used to determine which units are to be selected.

SIMPLE RANDOM SAMPLING….

- Estimates are easy to calculate.

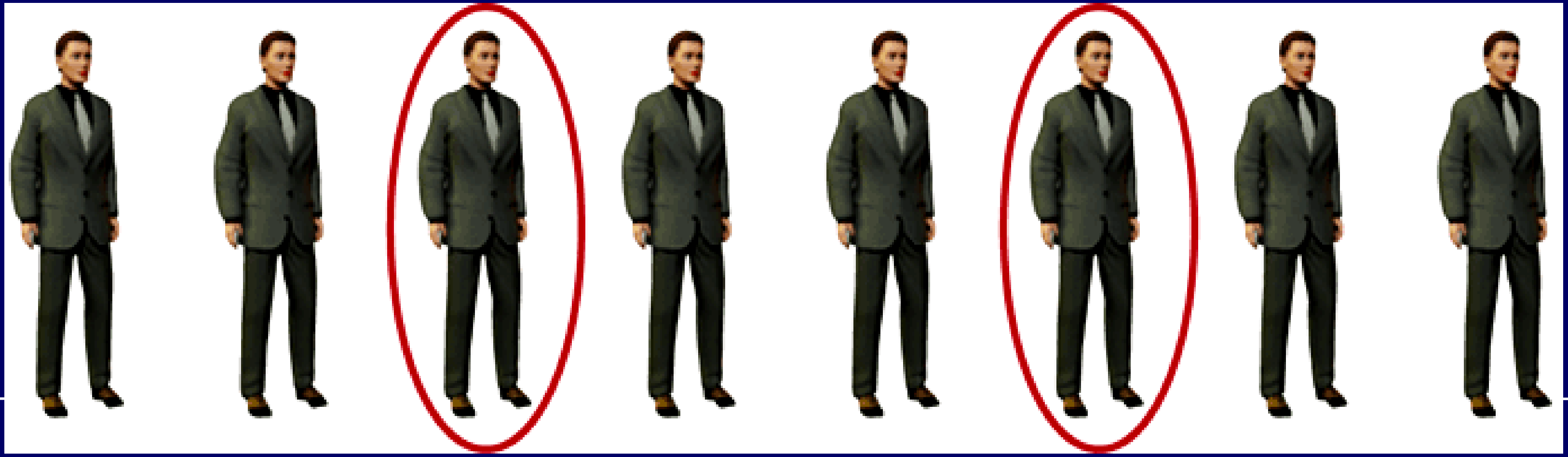
- **Disadvantages**
- If sampling frame large, this method impracticable.
- Minority subgroups of interest in population may not be present in sample in sufficient numbers for study.

SYSTEMATIC SAMPLING

- **Systematic sampling** relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list.
- Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards. In this case, *k*=(population size/sample size)
- It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list.
- *A* simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

SYSTEMATIC SAMPLING……

As described above, systematic sampling is a method, because all elements have the same probability of selection (in the example given, one in ten). It is *not* 'simple random sampling' because different subsets of the same size have different selection probabilities - e.g. the set {4,14,24,...,994} has a one-in-ten probability of selection, but the set {4,13,24,34,...} has zero probability of selection.



SYSTEMATIC SAMPLING……

- **ADVANTAGES:**
- Sample easy to select
- Suitable sampling frame can be identified easily
- Sample evenly spread over entire reference population
- **DISADVANTAGES:**
- Sample may be biased if hidden periodicity in population coincides with that of selection.
- Difficult to assess precision of estimate from one survey.

STRATIFIED SAMPLING

Where population embraces a number of distinct categories, the frame can be organized into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.

- Every unit in a stratum has same chance of being selected.
- Using same sampling fraction for all strata ensures proportionate representation in the sample.
- Adequate representation of minority subgroups of interest can be ensured by stratification & varying sampling fraction between strata as required.

STRATIFIED SAMPLING……

- Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata.

- **Drawbacks** to using stratified sampling.
- First, sampling frame of entire population has to be prepared separately for each stratum
- Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata.
- Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods

STRATIFIED SAMPLING……

Draw a sample from each stratum

CLUSTER SAMPLING

- Cluster sampling is an example of 'two-stage sampling' .
- First stage a sample of areas is chosen;
- Second stage a sample of respondents *within* those areas is selected.
- Population divided into clusters of homogeneous units, usually based on geographical contiguity.
- Sampling units are groups rather than individuals.
- A sample of such clusters is then selected.
- All units from the selected clusters are studied.

CLUSTER SAMPLING……

- Advantages :
- Cuts down on the cost of preparing a sampling frame.
- This can reduce travel and other administrative costs.
- Disadvantages: sampling error is higher for a simple random sample of same size.
- Often used to evaluate vaccination coverage

CLUSTER SAMPLING……

- **Identification of clusters**
  - List all cities, towns, villages & wards of cities with their population falling in target area under study.
  - Calculate cumulative population & divide by 30, this gives sampling interval.
  - Select a random no. less than or equal to sampling interval having same no. of digits. This forms 1^{st} cluster.
  - Random no.+ sampling interval = population of 2^{nd} cluster.
  - Second cluster + sampling interval = 4^{th} cluster.
  - Last or 30^{th} cluster = 29^{th} cluster + sampling interval

CLUSTER SAMPLING……

Two types of cluster sampling methods.

One-stage sampling. All of the elements within selected clusters are included in the sample.

Two-stage sampling. A subset of elements within selected clusters are randomly selected for inclusion in the sample.

Difference Between Strata and Clusters

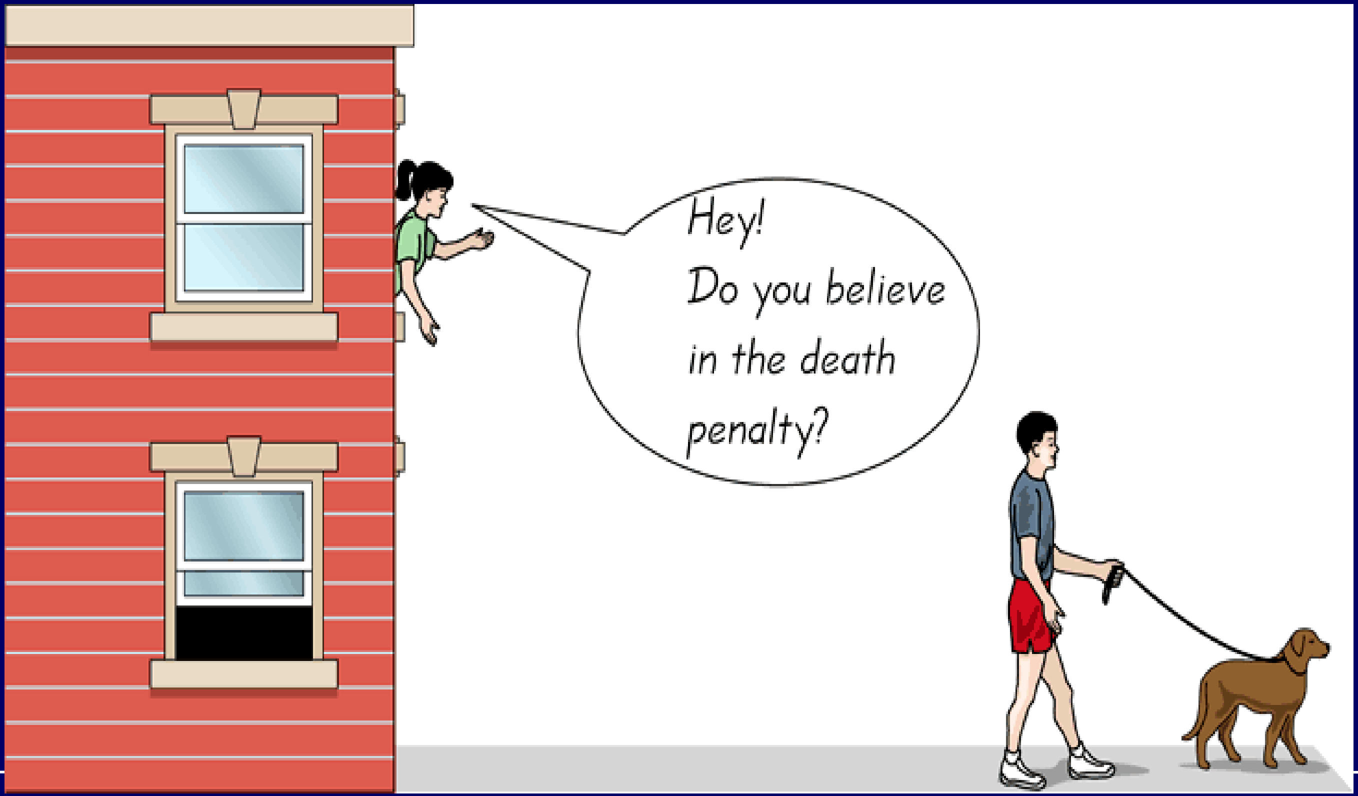
- Although strata and clusters are both non-overlapping subsets of the population, they differ in several ways.
- All strata are represented in the sample; but only a subset of clusters are in the sample.
- With stratified sampling, the best survey results occur when elements within strata are internally homogeneous. However, with cluster sampling, the best results occur when elements within clusters are internally heterogeneous

CONVENIENCE SAMPLING

- Sometimes known as **grab** or **opportunity sampling or accidental or haphazard sampling**.
- A type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, readily available and convenient.
- The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough.
- For example, if the interviewer was to conduct a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey was to be conducted at different times of day and several times per week.
- This type of sampling is most useful for pilot testing.
- In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample.

CONVENIENCE SAMPLING……

- Use results that are easy to get



Statistical Inference

The process of making guesses about the truth from a sample.

Sample statistics

Truth (not observable)

Population parameters

$$\mu = \frac{\sum_{i=1}^{N} x}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Sample (observation)

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X}_n)^2}{n-1}$$

*hat notation ^ is often used to indicate "estimate"

Make guesses about the whole population

Interval Estimation

- Population Mean: σ Known
- Population Mean: σ Unknown
- Determining the Sample Size
- Population Proportion

Margin of Error and the Interval Estimate

A point estimator cannot be expected to provide the exact value of the population parameter.

An interval estimate can be computed by adding and subtracting a margin of error to the point estimate.

Point Estimate +/– Margin of Error

The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.

Margin of Error and the Interval Estimate

The general form of an interval estimate of a population mean is

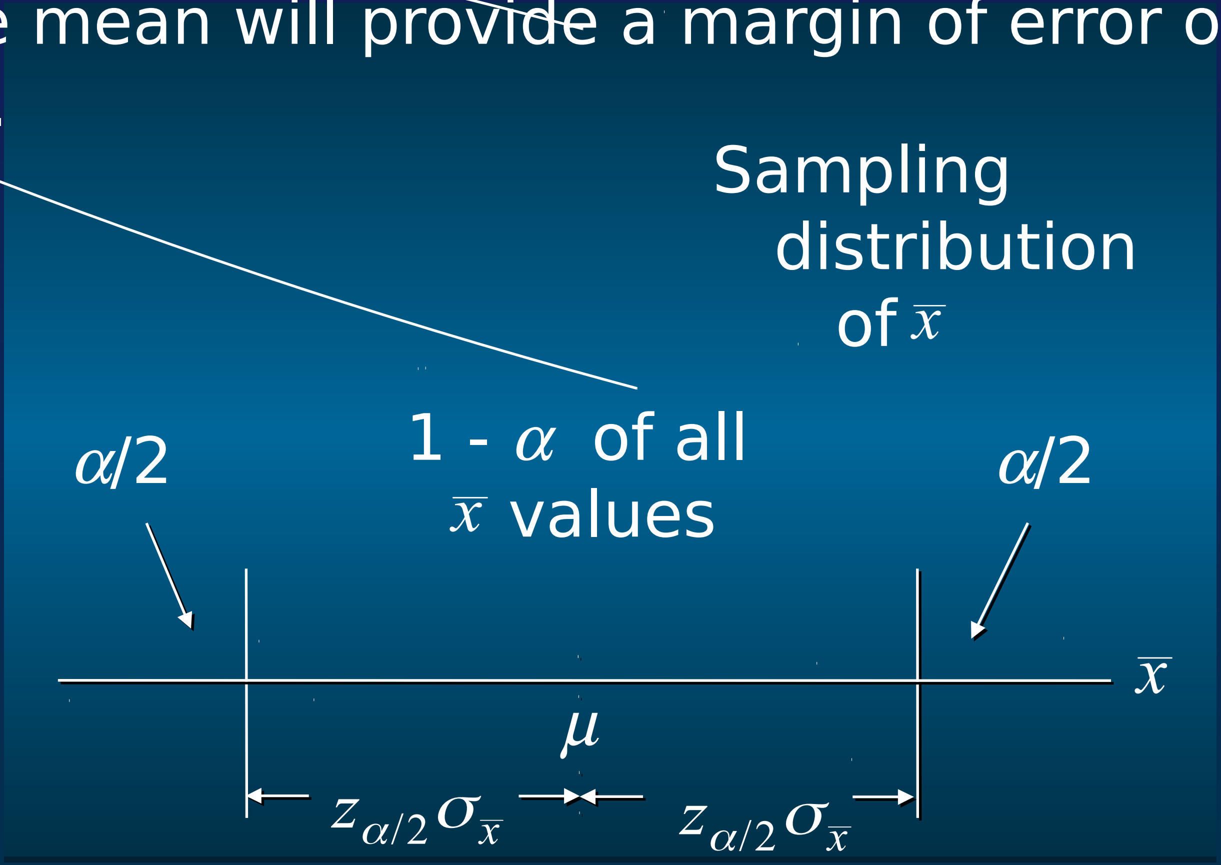
$$\bar{x} \pm \text{Margin of Error}$$

Interval Estimate of a Population Mean: σ Known

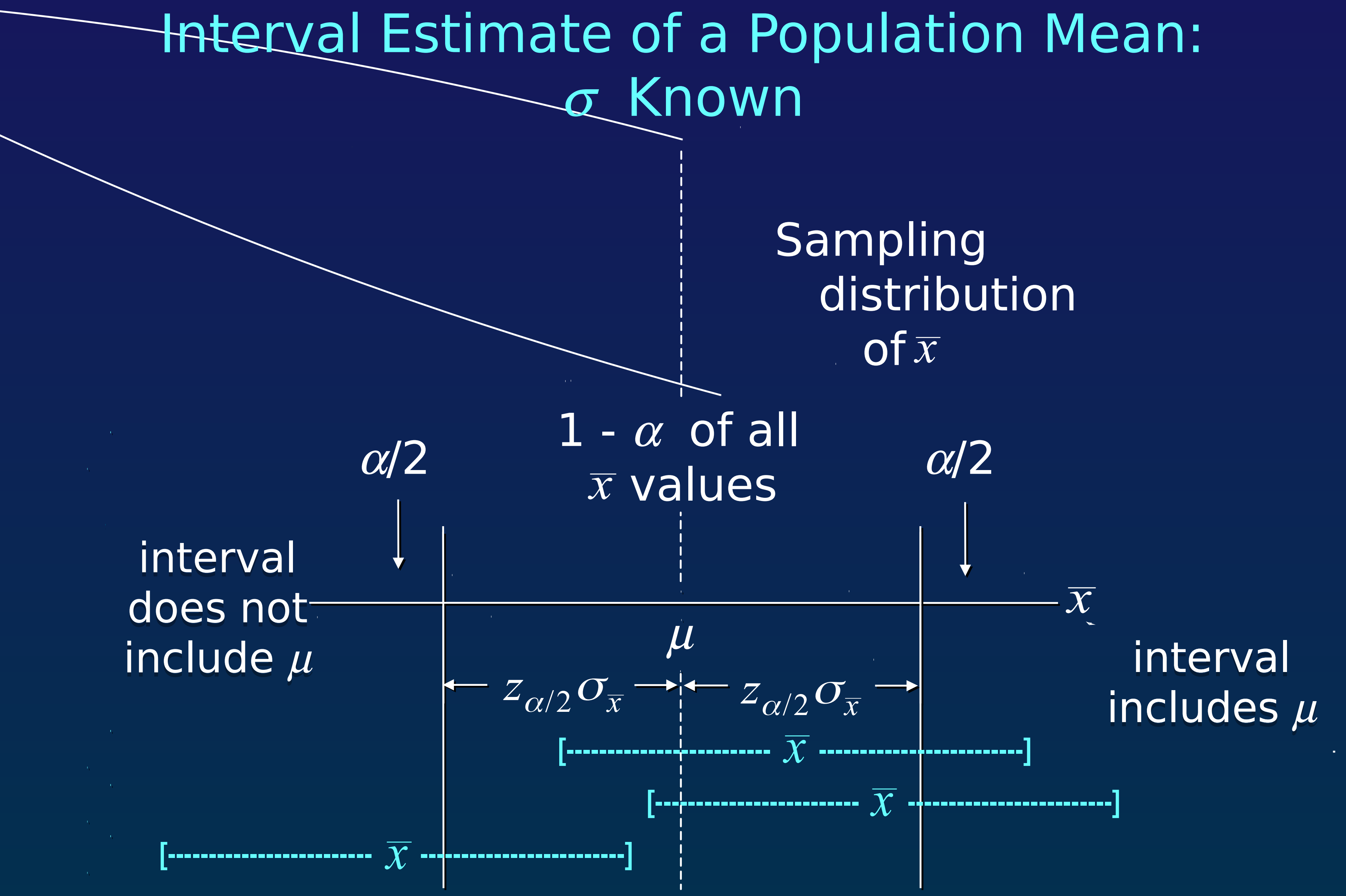
- In order to develop an interval estimate of a population mean, the margin of error must be computed using either:
  - the population standard deviation σ, or
  - the sample standard deviation s
- σ is rarely known exactly, but often a good estimate can be obtained based on historical data or other information.
- We refer to such cases as the <u>σ known</u> case.

Interval Estimate of a Population Mean: σ Known

There is a $1 - \alpha$ probability that the value of a sample mean will provide a margin of error of $z_{\alpha/2}\sigma_{\bar{x}}$ or less.



Interval Estimate of a Population Mean: σ Known



Interval Estimate of a Population Mean: σ Known

- Interval Estimate of μ

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where: $\bar{x}$ is the sample mean

$1 - \alpha$ is the confidence coefficient

$z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution

σ is the population standard deviation

n is the sample size

Interval Estimate of a Population Mean: σ Known

- Interval Estimate of μ

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where: $\bar{x}$ is the sample mean

$1 - \alpha$ is the confidence coefficient

$z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution

σ is the population standard deviation

n is the sample size

Meaning of Confidence

Because 90% of all the intervals constructed using $\bar{x} \pm 1.645\sigma_{\bar{x}}$ will contain the population mean, we say we are 90% confident that the interval $\bar{x} \pm 1.645\sigma_{\bar{x}}$ includes the population mean μ.

We say that this interval has been established at the 90% confidence level.

The value .90 is referred to as the confidence coefficient.

Interval Estimate of a Population Mean: σ Known

- Example: Discount Sounds

Discount Sounds has 260 retail outlets throughout the United States. The firm is evaluating a potential location for a new outlet, based in part, on the mean annual income of the individuals in the marketing area of the new location.

A sample of size n = 36 was taken; the sample mean income is \$41,100. The population is not believed to be highly skewed. The population standard deviation is estimated to be \$4,500, and the confidence coefficient to be used in the interval estimate is .95.

Interval Estimate of a Population Mean: σ Known

- Example: Discount Sounds

95% of the sample means that can be observed are within $\pm$ $1.96\sigma_{\bar{x}}$ of the population mean μ.

The margin of error is:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96\left(\frac{4,500}{\sqrt{36}}\right) = 1,470$$

Thus, at 95% confidence,
the margin of error is $1,470.

Interval Estimate of a Population Mean: σ Known

- Example: Discount Sounds

Interval estimate of μ is:

$41,100 $\underline{+}$ $1,470

or

$39,630 to $42,570

We are <u>95% confident</u> that the interval contains the population mean.

Interval Estimate of a Population Mean: σ Known

- Adequate Sample Size

In most applications, a sample size of $n = 30$ is adequate.

If the population distribution is highly skewed or contains outliers, a sample size of 50 or more is recommended.

Interval Estimate of a Population Mean: σ Known

- Adequate Sample Size (continued)

If the population is not normally distributed but is roughly symmetric, a sample size as small as 15 will suffice.

If the population is believed to be at least approximately normal, a sample size of less than 15 can be used.

Interval Estimate of a Population Mean: σ Unknown

- If an estimate of the population standard deviation σ cannot be developed prior to sampling, we use the sample standard deviation s to estimate σ.
- This is the <u>σ unknown</u> case.
- In this case, the interval estimate for μ is based on the t distribution.
- (We'll assume for now that the population is normally distributed.)

t Distribution

William Gosset, writing under the name "Student", is the founder of the t distribution.

Gosset was an Oxford graduate in mathematics and worked for the Guinness Brewery in Dublin.

He developed the t distribution while working on small-scale materials and temperature experiments.

t Distribution

The <u>*t* distribution</u> is a family of similar probability distributions.

A specific *t* distribution depends on a parameter known as the <u>degrees of freedom</u>.

Degrees of freedom refer to the number of independent pieces of information that go into the computation of *s*.

t Distribution

A *t* distribution with more degrees of freedom has less dispersion.

As the degrees of freedom increases, the difference between the *t* distribution and the standard normal probability distribution becomes smaller and smaller.

t Distribution

For more than 100 degrees of freedom, the standard normal z value provides a good approximation to the t value.

The standard normal z values can be found in the infinite degrees (∞) row of the t distribution table.

t Distribution

Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
.	.	.	.	.	.	.
50	.849	1.299	1.676	2.009	2.403	2.678
60	.848	1.296	1.671	2.000	2.390	2.660
80	.846	1.292	1.664	1.990	2.374	2.639
100	.845	1.290	1.660	1.984	2.364	2.626
∞	.842	1.282	1.645	1.960	2.326	2.576

Standard normal
z values

Interval Estimate of a Population Mean: σ Unknown

- Interval Estimate

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where: $1 - \alpha$ = the confidence coefficient

$t_{\alpha/2}$ = the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with n - 1 degrees of freedom

s = the sample standard deviation

Interval Estimate of a Population Mean: σ Unknown

- Example: Apartment Rents

A reporter for a student newspaper is writing an article on the cost of off-campus housing. A sample of 16 one-bedroom apartments within a half-mile of campus resulted in a sample mean of $750 per month and a sample standard deviation of $55.

Let us provide a 95% confidence interval estimate of the mean rent per month for the population of one-bedroom efficiency apartments within a half-mile of campus. We will assume this population to be normally distributed.

Interval Estimate of a Population Mean: σ Unknown

At 95% confidence, $\alpha = .05$, and $\alpha/2 = .025$.

$t_{.025}$ is based on $n - 1 = 16 - 1 = 15$ degrees of freedom.

In the t distribution table we see that $t_{.025} = 2.131$.

Degrees	Area in Upper Tail					
of Freedom	**.20**	**.100**	**.050**	**.025**	**.010**	**.005**
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.520	2.878
19	.861	1.328	1.729	2.093	2.539	2.861
.	.	.	.	.	.	.

Interval Estimate of a Population Mean: σ Unknown

- Interval Estimate

$$\bar{x} \pm t_{.025} \frac{s}{\sqrt{n}}$$

Margin of Error

$$750 \pm 2.131 \frac{55}{\sqrt{16}} = 750 \pm \underline{29.30}$$

We are 95% confident that the mean rent per month for the population of one-bedroom apartments within a half-mile of campus is between $720.70 and $779.30.

Interval Estimate of a Population Mean: σ Unknown

- Adequate Sample Size

In most applications, a sample size of n = 30 is adequate when using the expression $\bar{x} \pm t_{\alpha/2}\, s/\sqrt{n}$ to develop an interval estimate of a population mean.

If the population distribution is highly skewed or contains outliers, a sample size of 50 or more is recommended.

Interval Estimate of a Population Mean: σ Unknown

- Adequate Sample Size (continued)

If the population is not normally distributed but is roughly symmetric, a sample size as small as 15 will suffice.

If the population is believed to be at least approximately normal, a sample size of less than 15 can be used.

Sample Size for an Interval Estimate of a Population Mean

Let E = the desired margin of error.

E is the amount added to and subtracted from the point estimate to obtain an interval estimate.

If a desired margin of error is selected prior to sampling, the sample size necessary to satisfy the margin of error can be determined.

Sample Size for an Interval Estimate of a Population Mean

- Margin of Error

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Necessary Sample Size

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

Sample Size for an Interval Estimate of a Population Mean

The Necessary Sample Size equation requires a value for the population standard deviation σ.

If σ is unknown, a preliminary or planning value for σ can be used in the equation.

1. Use the estimate of the population standard deviation computed in a previous study.
2. Use a pilot study to select a preliminary study and use the sample standard deviation from the study.
3. Use judgment or a “best guess” for the value of σ.

Sample Size for an Interval Estimate of a Population Mean

- Example: Discount Sounds

Recall that Discount Sounds is evaluating a potential location for a new retail outlet, based in part, on the mean annual income of the individuals in the marketing area of the new location.

Suppose that Discount Sounds' management team wants an estimate of the population mean such that there is a .95 probability that the sampling error is $500 or less.

How large a sample size is needed to meet the required precision?

Sample Size for an Interval Estimate of a Population Mean

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 500$$

At 95% confidence, $z_{.025}$ = 1.96. Recall that σ = 4,500.

$$n = \frac{(1.96)^2(4,500)^2}{(500)^2} = 311.17 = 312$$

A sample of size 312 is needed to reach a desired precision of $\pm$ \$500 at 95% confidence.

Interval Estimate of a Population Proportion

The general form of an interval estimate of a population proportion is

$$\bar{p} \pm \text{Margin of Error}$$

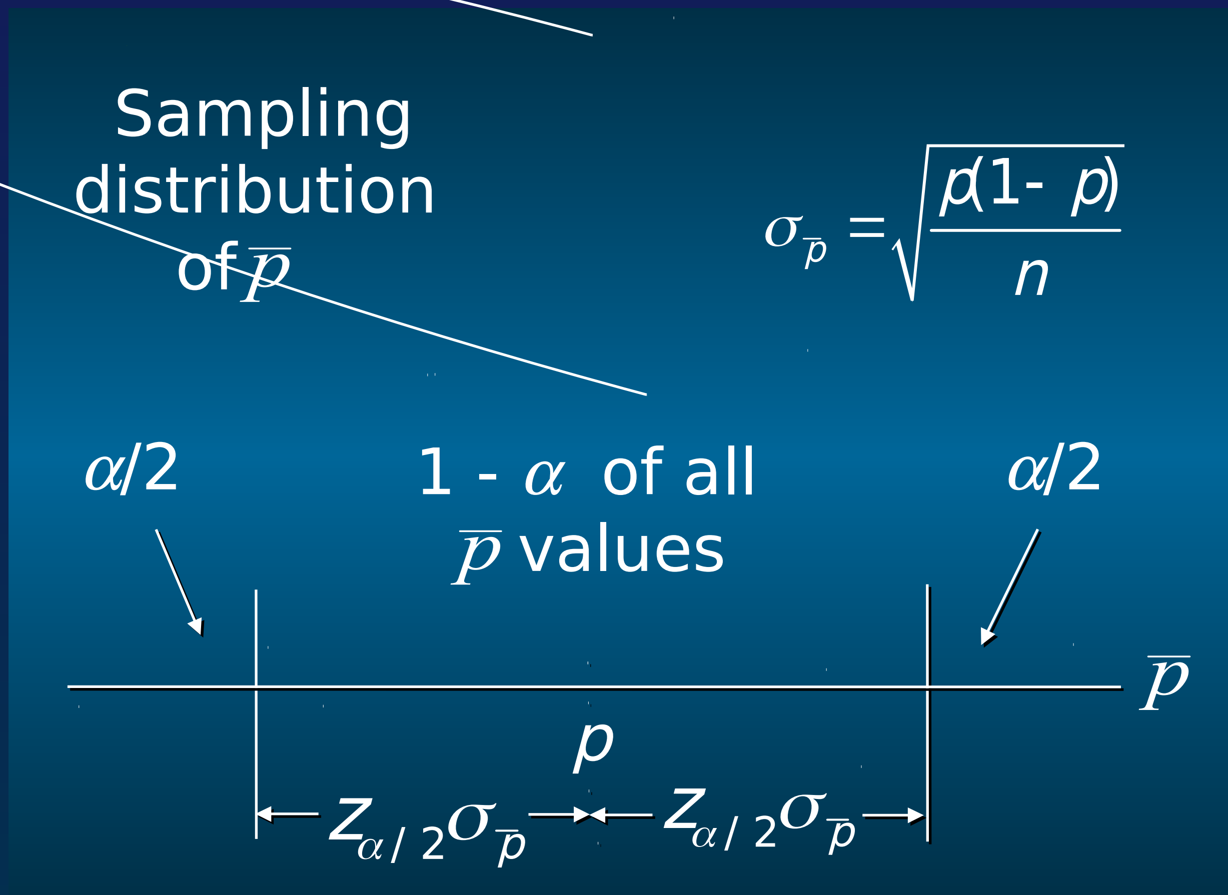
Interval Estimate of a Population Proportion

The sampling distribution of $\bar{p}$ plays a key role in computing the margin of error for this interval estimate.

The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1 - p) \geq 5$.

Interval Estimate of a Population Proportion

- Normal Approximation of Sampling Distribution of $\bar{p}$



Interval Estimate of a Population Proportion

- Interval Estimate

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where: $1-\alpha$ is the confidence coefficient

$z_{\alpha/2}$ is the z value providing an area of $\alpha/2$ in the upper tail of the standard normal probability distribution

$\bar{p}$ is the sample proportion

Interval Estimate of a Population Proportion

- Example: Political Science, Inc.

Political Science, Inc. (PSI) specializes in voter polls and surveys designed to keep political office seekers informed of their position in a race.

Using telephone surveys, PSI interviewers ask registered voters who they would vote for if the election were held that day.

Interval Estimate of a Population Proportion

- Example: Political Science, Inc.

  In a current election campaign, PSI has just found that 220 registered voters, out of 500 contacted, favor a particular candidate. PSI wants to develop a 95% confidence interval estimate for the proportion of the population of registered voters that favor the candidate.

Interval Estimate of a Population Proportion

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where: $n = 500$, $\bar{p} = 220/500 = .44$, $z_{\alpha/2} = 1.96$

$$.44 \pm 1.96 \sqrt{\frac{.44(1-.44)}{500}} = .44 + .0435$$

PSI is 95% confident that the proportion of all voters that favor the candidate is between .3965 and .4835.

Sample Size for an Interval Estimate of a Population Proportion

- Margin of Error

$$E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Solving for the necessary sample size, we get

$$n = \frac{(z_{\alpha/2})^2\,\bar{p}(1-\bar{p})}{E^2}$$

However, $\bar{p}$ will not be known until after we have selected the sample. We will use the planning value p^* for $\bar{p}$.

Sample Size for an Interval Estimate of a Population Proportion

- Necessary Sample Size

$$n = \frac{(z_{\alpha/2})^2 p^*(1 - p^*)}{E^2}$$

The planning value p^* can be chosen by:

1. Using the sample proportion from a previous sample of the same or similar units, or
2. Selecting a preliminary sample and using the
   sample proportion from this sample.
3. Use judgment or a “best guess” for a p^* value.
4. Otherwise, use .50 as the p^* value.

Sample Size for an Interval Estimate of a Population Proportion

- Example: Political Science, Inc.

Suppose that PSI would like a .99 probability that the sample proportion is within + .03 of the population proportion. How large a sample size is needed to meet the

required precision? (A previous sample of similar units yielded .44 for the sample proportion.)

Sample Size for an Interval Estimate of a Population Proportion

$$z_{\alpha/2}\sqrt{\frac{p^*(1-p^*)}{n}} = .03$$

At 99% confidence, $z_{.005}$ = 2.576. Recall that p^* = .44.

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2} = \frac{(2.576)^2(.44)(.56)}{(.03)^2} \cong 1817$$

A sample of size 1817 is needed to reach a desired precision of $\pm$ .03 at 99% confidence.

Z-score Review

- A sample mean (X-bar) approximates a population mean (μ)
- The standard error provides a measure of
  - how well a sample mean approximates the population mean
  - determines how much difference between X-bar and μ is reasonable to expect just by chance
- The z-score is a statistic used to quantify this inference

$$z = \frac{\overline{X} - \mu}{\sigma}$$

- obtained difference between data and hypothesis/standard distance expected by chance

What is the t statistic?

- “Cousin” of the z statistic that does not require the population mean (μ) or variance (σ^2)to be known
- Can be used to test hypotheses about a completely unknown population (when the only information about the population comes from the sample)
- Required: a sample and a reasonable hypothesis about the population mean (μ)
- Can be used with one sample or to compare two samples

From Z to *t*…

$$Z = \frac{\overline{X} - \mu_{hyp}}{\sigma_{\overline{X}}}$$

$$t = \frac{\overline{X} - \mu_{hyp}}{s_{\overline{X}}}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Note lowercase "s".

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}}$$

$$\sigma = \sqrt{\frac{N\Sigma X^2 - (\Sigma X)^2}{N^2}}$$

$$s = \sqrt{\frac{n\Sigma X^2 - (\Sigma X)^2}{n(n - 1)}}$$

Thank You…..!!!!!

Any Questions ?????