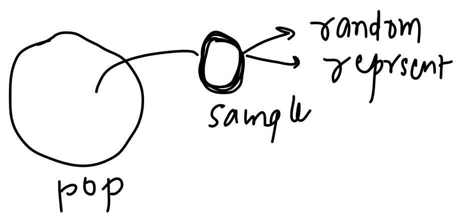
Some Terms

30 March 2023 07:09



Population Vs Sample

Population: A population is the entire group or set of individuals, objects, or events that a researcher wants to study or draw conclusions about. It can be people, animals, plants, or even inanimate objects, depending on the context of the study. The population usually represents the complete set of possible data points or observations.

Sample: A sample is a subset of the population that is selected for study. It is a smaller group that is intended to be representative of the larger population. Researchers collect data from the sample and use it to make inferences about the population as a whole. Since it is often impractical or impossible to collect data from every member of a population, samples are used as an efficient and cost-effective way to gather information.

pop → prop, var, mean, std } → parameter

Parameter Vs ~~Estima~~te → statistics

Parameter: A parameter is a numerical value that describes a characteristic of a population. Parameters are usually denoted using Greek letters, such as μ (mu) for the population mean or σ (sigma) for the population standard deviation. Since it is often difficult or impossible to obtain data from an entire population, parameters are usually unknown and must be estimated based on available sample data.

Greek word হলে, বুঝবো Population.

$\bar{x} \rightarrow \mu$

sample → mean ($\bar{x}$), std(s), var } → stastic

Statistic: A statistic is a numerical value that describes a characteristic of a sample, which is a subset of the population. By using statistics calculated from a representative sample, researchers can make inferences about the unknown respective parameter of the population. Common statistics include the sample mean (denoted by x̄, pronounced "x-bar"), the sample median, and the sample standard deviation (denoted by s).

→ $\bar{x}$ হলে sample থেকে derive করা।

sample → population

Inferential Statistics

Inferential statistics is a branch of statistics that focuses on making predictions, estimations, or generalizations about a larger population based on a sample of data taken from that population. It involves the use of probability theory to make inferences and draw conclusions about the characteristics of a population by analysing a smaller subset or sample.

The key idea behind inferential statistics is that it is often impractical or impossible to collect data from every member of a population, so instead, we use a representative sample to make inferences about the entire group. Inferential statistical techniques include hypothesis testing, confidence intervals, and regression analysis, among others. ***

inferential stat.
① আমরা ② prediction or estimation করি ২টা population এর উপর। based on sample.

These methods help researchers answer questions like:

a. Is there a significant difference between two groups?
b. Can we predict the outcome of a variable based on the values of other variables?
c. What is the relationship between two or more variables?

Inferential statistics are widely used in various fields, such as economics, social sciences, medicine, and natural sciences, to make informed decisions and guide policy based on limited data.

Point Estimate

30 March 2023 07:19

→ (A single value calculate from a sample)

A point estimate is a [single value] calculated from a sample, that serves as the best guess or approximation for an unknown population parameter, such as the mean or standard deviation. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.

mean

Subs → 77K → avg age → 28

live → 100 → $\bar{X}$ → sample mean

11 live class

[point estimate]

10 live class → 100 → point estimate

10 classes

$\bar{X}_1, \bar{X}_2, \bar{X}_3 \ldots \bar{X}_{10}$ → $\bar{X}_{mean}$ → point estimate

এখানে CLT ব্যবহার করে $\bar{x}$mean calculate করা হয়েছে যেটাও point estimation.

28 → YT subs → avg age

reliable → NO

35 → 28

point estimates

range

calculate

Confidence interval

MS Dhoni

1) exact → 25 → 25 → 100000 (1 cr)

2) ±10 → 7500 (75 lac) ✓

3) ±20 → 500 (50 lac) ✓

Problem in point estimation and introduction to confident interval:

উপরের, example এ, আমরা বাজি ধরেছি Dhoni, ২৫ রান করলে 1 কোটি টাকা, (25±10) হলে 75 লাখ, (25±20) 50 লাখ। এখানে, Dhoni আসলে ২৫ রানই করবে এটা confident interval। কিন্তু, আমাদের এই prediction successful হওয়ার possibility কম। কিন্তু, আমি যদি এটাকে range আকারে, (point estimation ± Error) আকারে প্রকাশ করি, তাহলে successful হওয়ার possibility বেড়ে যাবে। আর, point estimation-কে range আকারে প্রকাশ করাই হচ্ছে [Confidence Interval].

Confidence Interval

30 March 2023 07:18

μ, σ $\bar{x}$ [25, 32] 95% confident

Confidence interval, in simple words, is a range of values within which we expect a particular population parameter, like a mean, to fall. It's a way to express the uncertainty around an estimate obtained from a sample of data.

Confidence level, usually expressed as a percentage like 95%, indicates how sure we are that the true value lies within the interval.

95% 25 ± 4 [21, 29]

Confidence Interval = Point Estimate ± Margin of Error

Ways to calculate CI:

কীভাবে CI বের করবো

Z procedure

Pop → std available (σ)

Population এর std জানা থাকলে

t procedure

[pop - std (σ) X

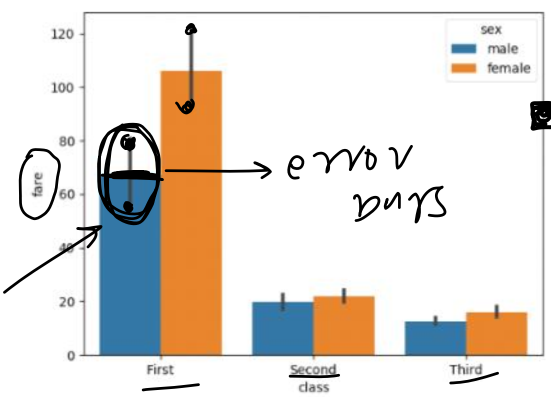
Population এর std না জানা থাকলে।

Confidence Interval is created for Parameters and not statistics. Statistics help us get the confidence interval for a parameter.

Confidence interval population এর term এ হয়। Not statistic (sample).

Examples of CT usage

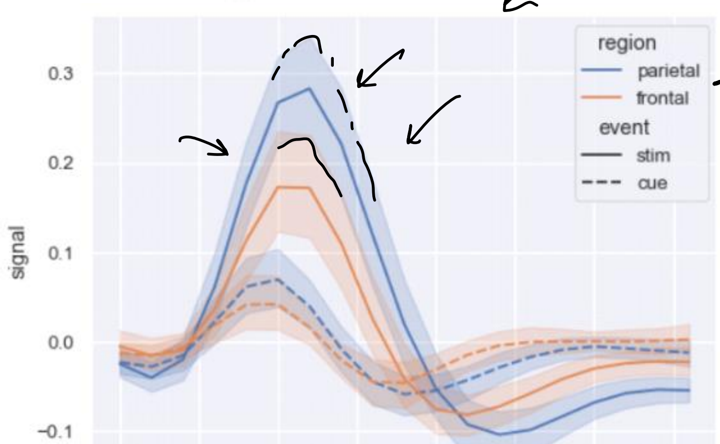
seaborn → bar plot



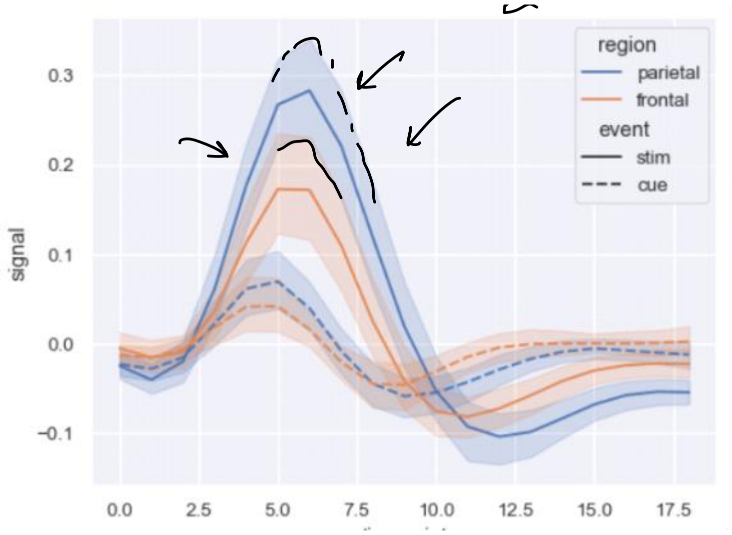
error bars

(CI)

seaborn এ CI বার গুলোকে error bars বলে, mainly এটা হচ্ছে CI বা confidence interval।



এখানে, shade portion জোন রাখা করে। যা CI নির্দেশ করে।


region
parietal
frontal
event
stim
cue
signal
0.3
0.2
0.1
0.0
−0.1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

Confidence Interval (Sigma Known)

Z procedure

30 March 2023 07:13

(σ) pop std available

pop → sample random

Assumptions

1. **Random sampling**: The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
2. **Known population standard deviation**: The population standard deviation (σ) must be known or accurately estimated. In practice, the population standard deviation is often unknown, and the sample standard deviation (s) is used as an estimate. However, if the sample size is large enough, the sample standard deviation can provide a reasonably accurate approximation.
3. **Normal distribution or large sample size**: The Z-procedure assumes that the underlying population is normally distributed. However, if the population distribution is not normal, the Central Limit Theorem can be applied when the sample size is large (usually, sample size n ≥ 30 is considered large enough). According to the Central Limit Theorem, the sampling distribution of the sample mean will approach a normal distribution as the sample size increases, regardless of the shape of the population distribution.

Distribution Normal হতে হবে। যদি না হয় তাহলে CLT ব্যবহার করে

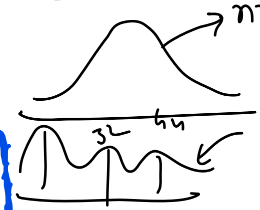
Sample size ≤ 30 — 20 → z procedure

YT – pop → std (σ)

distribution → normal

77K subs → age

normal dist



clt

A (1 - alpha)*100% Confidence Interval for mu:

$\sigma = 15$

YT → campusx → 77K → 28 ± 14 → [16, 42] ← confidence interval

Confidence level → 95%

formula CI using Z procedure

$$CI = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

1) Intuition

2) $Z_{\alpha/2}$

where

$Z \to ?$

$(1-\alpha) \to$ confidence level

$(1-\alpha) \to 95\%$

$\sigma \to$ std pop

$n \to$ sample size → 100

YouTube example

~~কোনো~~ YT channel এ subscribe এদের avg. Age ~~কত~~ ?

Intuition

point estimate ($\bar{x}$) → CLT

10 live class নিয়ে 50 জন এর age নিয়ে CLT ব্যবহার করলাম।

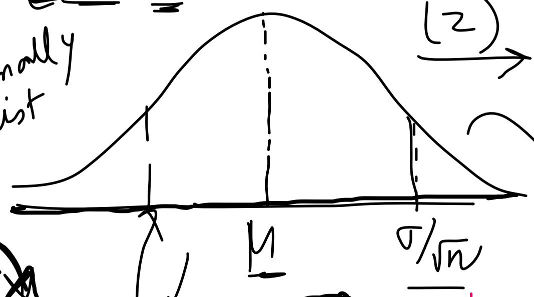
point estimate $(\bar{x}) \rightarrow$ CLT

10 live class $\rightarrow$ 50 $\rightarrow$ avg class

$\bar{x}_1, \bar{x}_2 \ldots \ldots \bar{x}_{10} \rightarrow$ sampling dist of sample mean

normally dist

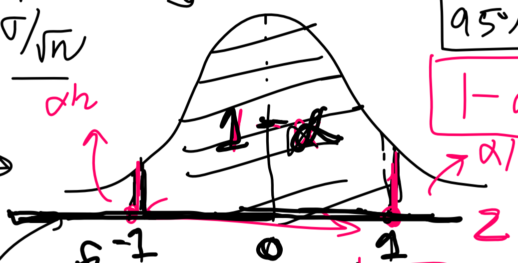
(z) $\rightarrow N(0,1)$



standard normal variate $\mu=0, \sigma=1$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

(z)



95%

$1-\alpha$ prob $\rightarrow$ 95%

Z lie

z: $-z_{\alpha/2}$, $z_{\alpha/2}$

$-z_{2.5} \longleftrightarrow z_{2.5}$

95%

$95\% = (1-\alpha)$ sure

হলে, $\alpha = 5\%$

$\alpha = 5$

formula

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

95

$z_{\alpha/2}$ হচ্ছে z-values

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

Point Estimator আমরা μ বের করে তারপর Confidence Interval বের করবো। তারপর $\mu \rightarrow$ CI আলাদা করবো।

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{x}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$P\left(-\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < \ldots + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

fixed

95%

$$\boxed{P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \boxed{\mu} < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \boxed{1-\alpha}}$$

$1-\alpha$ 0.95 $\alpha = 0.50/2$

$\bar{x} \rightarrow$ sample

simplified equation $\rightarrow$ CI

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

confidence 0.250

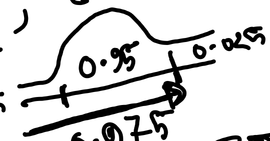
যদি $(1-\alpha) \rightarrow 95\%$

তাহলে, $\alpha/2 = \frac{5\%}{2} = 0.025$

$$\mu = \bar{x} \pm z_{0.0250}\frac{\sigma}{\sqrt{n}}$$

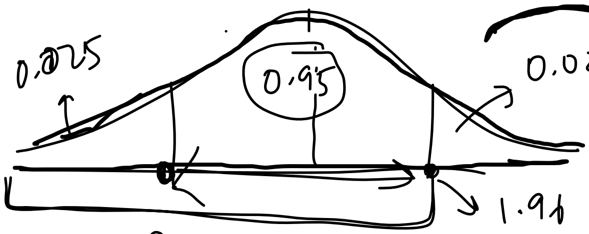
z

এর পর, distribution থেকে



950
025
975

এর জন্য z-table থেকে confidence value বের করতে হবে, $z_{\alpha/2}$ এর মান।



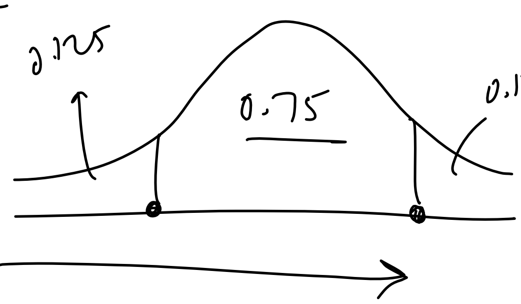
range $\rightarrow$ confidence interval

$$\boxed{\mu = \bar{x} \pm \boxed{1.96}\frac{\sigma}{\sqrt{n}}} \rightarrow \text{CI with a confidence level} - 95\%$$

$z_{\alpha/2} \rightarrow 1.96$

50% 75% 99%

0.87 $z_{\alpha/2}$



$z \rightarrow 1.13$

Interpreting Confidence Interval

30 March 2023 08:33

pop fixed

45 → [14 – 42]

A confidence interval is a range of values within which a population parameter, such as the population mean, is estimated to lie with a certain level of confidence. The confidence interval provides an indication of the precision and uncertainty associated with the estimate. To interpret the confidence interval values, consider the following points:

95% → [18, 42]

77K → 100 times

random samples (50)

age → 1

age → 2

3

100 times

95 times

1. **Confidence level**: The confidence level (commonly set at 90%, 95%, or 99%) represents the probability that the confidence interval will contain the true population parameter if the sampling and estimation process were repeated multiple times. For example, a 95% confidence interval means that if you were to draw 100 different samples from the population and calculate the confidence interval for each, approximately 95 of those intervals would contain the true population parameter.
2. **Interval range**: The width of the confidence interval gives an indication of the precision of the estimate. A narrower confidence interval suggests a more precise estimate of the population parameter, while a wider interval indicates greater uncertainty. The width of the interval depends on the sample size, variability in the data, and the desired level of confidence.
3. **Interpretation**: To interpret the confidence interval values, you can say that you are "X% confident that the true population parameter lies within the range (lower limit, upper limit)." Keep in mind that this statement is about the interval, not the specific point estimate, and it refers to the confidence level you chose when constructing the interval.

What is the trade-off

Confident level | width

95% → width → ✓

100% → [0 200]

95% → 150 – 50

1% ← 25 →

[←→]

100% [−∞, ∞]

Cricket এ একজন batsman (0-200) run করার possibility বেশি থাকে। কারণ, কম্পুকেই এমন হয়। Theoretically, এই জন্য Confidence level 100% দিলে width (−∞,∞) হয়। এবং সবচেয়ে optimum value 95%

Factors Affecting Margin of Error

30 March 2023 07:15

1. Confidence Level (1-alpha)
2. Sample Size
3. Population Standard Deviation

Margin Error

$\dfrac{\text{upper} - \text{lower}}{2} \rightarrow$ marg.

CI = point estimate $\pm$ margin of error

$= \boxed{\bar{x}} \pm Z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

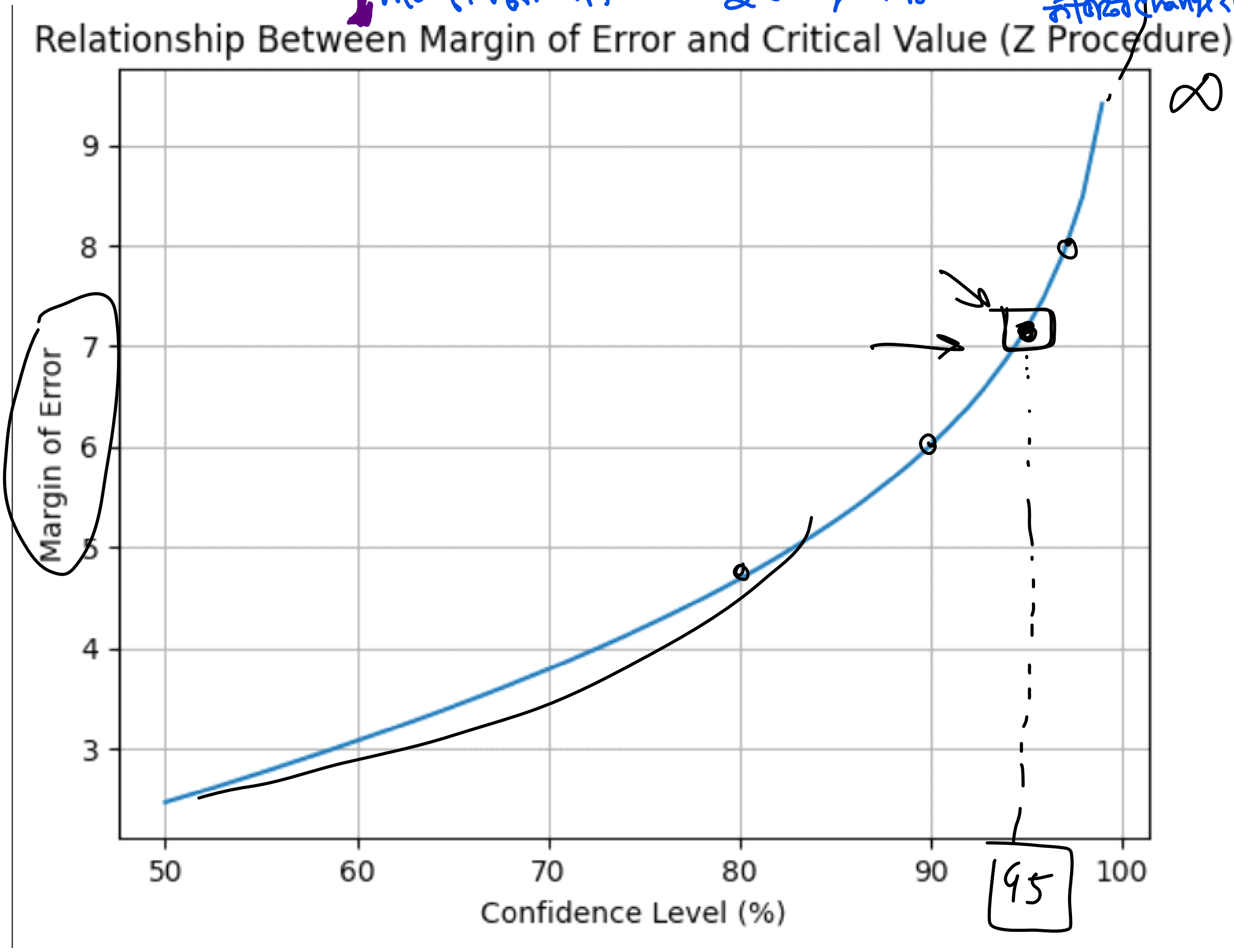
sample mean

(1) critical value

$\sigma \rightarrow$ pop std (2)

$\sqrt{n} \rightarrow$ sample size (3)

(i) critical value (ii) pop. std (iii) sample size এই 3 টার উপার CI (Z-Procedure) depend করে। নিচে দেখবো যদি ১ টা কিছু change করি তাহলে CI এর value কীভাবে change হয়



বাস্তবে সমস্যা হচ্ছে যে, z-Procedure এর জন্য আমাদের Population এর std জানা থাকা দরকার। কিন্তু, বাস্তবে সেটা পাওয়া যায় না। তাই আমরা, t-procedure ব্যবহার করি।

Confidence Interval (Sigma not known)

30 March 2023 07:15

$\boxed{\sigma} \rightarrow$ t-procedure

Using the t procedure → Assumption must be follow 100%

Assumptions

1. **Random sampling**: The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.
2. **Sample standard deviation**: The population standard deviation (σ) is unknown, and the sample standard deviation (s) is used as an estimate. The t-distribution is specifically designed to account for the additional uncertainty introduced by using the sample standard deviation instead of the population standard deviation.
3. **Approximately normal distribution**: The t-procedure assumes that the underlying population is approximately normally distributed, or the sample size is large enough for the Central Limit Theorem to apply. If the population distribution is heavily skewed or has extreme outliers, the t-procedure may not be accurate, and non-parametric methods should be considered.
4. **Independent observations**: The observations in the sample should be independent of each other. In other words, the value of one observation should not influence the value of another observation. This is particularly important when working with time series data or data with inherent dependencies.

majboori

আমাদের কাছে Population এর std নেই কিন্তু, sample এর আছে তো

sample > 30 → if it's not normal

normal → small sample size

sam

CT

$t_{\alpha/2} >> z_{\alpha/2}$

$t_{\alpha/2} \simeq z_{\alpha/2}$

Z-table

$$CI = \bar{X} \pm z_{\alpha/2} \frac{\boxed{\sigma}}{\sqrt{n}} \qquad \sigma \rightarrow s$$

complexity

Pop mean

$$CI = \bar{X} + \boxed{z_{\alpha/2}} \frac{s}{\sqrt{n}} \rightarrow CI = \bar{X} + \boxed{t_{\alpha/2}} \frac{s}{\sqrt{n}}$$

df > 30

95% $z_{\alpha/2} \rightarrow 1.96$ $t_{\alpha/2} = 2.0008$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow \text{standard normal dist}$$

$$z' \quad \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

student's t distribution

parameter

df → $n-1$

$n-1$

49 30

theoritical dist

student's t dis

$df \rightarrow \infty \quad (t \rightarrow N)$

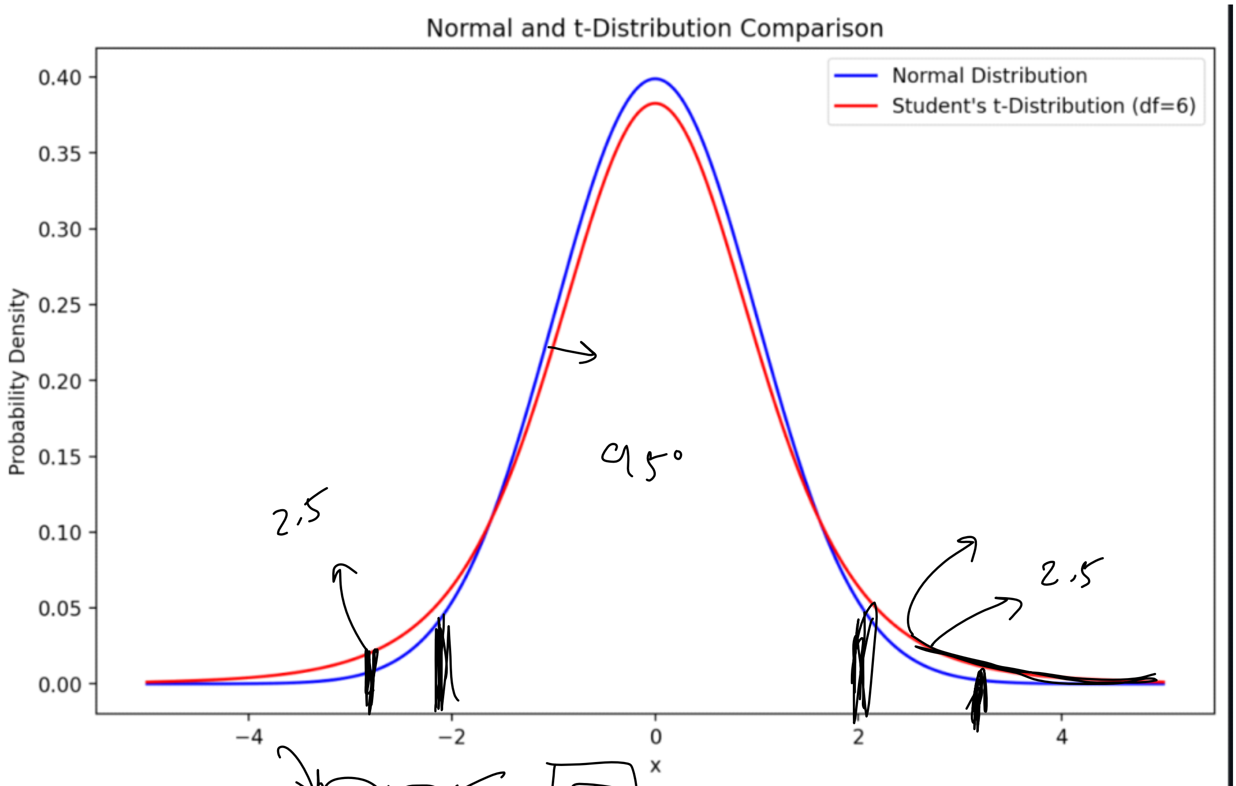
μ, σ

z-Procedure এ আমরা standard normal variate calculate করেছিলাম। কিন্তু t-তে, এ আমাদের কাছে sample এর std আছে এর জন্য

$$z' = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

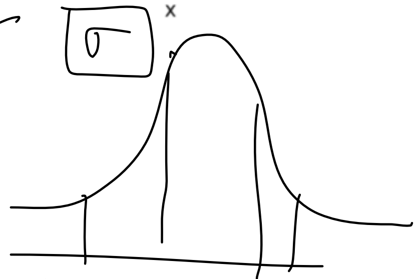
calculate করি, আর এর জন্য আমরা যে, distribution পাবো সেইটাকে student's t-distribution বলি।

Comparasion between Normal & t-distribution



$$CI = \bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$$

$\boxed{\sigma}$



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

$$\boxed{CI = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}} \longleftrightarrow \boxed{CI = \bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

CLT

$\bar{x}_1$ s_1

$\bar{x}_2$ s_2

s_3

⋮

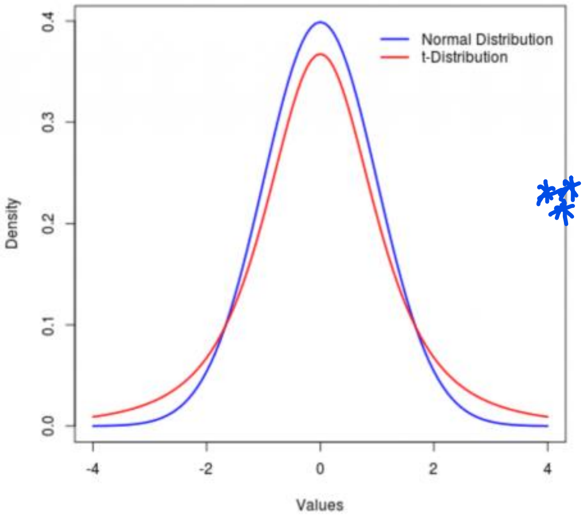
s_{10}

$\rightarrow$ s_{avg}

→ T' distribution এর জন্য t-table আছে, সেখান থেকে $t_{\alpha/2}$ এর মান বের করতে হয়।

Student's T Distribution

30 March 2023 07:16



Student's t-distribution, or simply the t-distribution, is a probability distribution that arises when estimating the mean of a normally distributed population when the sample size is small and the population standard deviation is unknown. It was introduced by William Sealy Gosset, who published under the pseudonym "Student."

** The t-distribution is similar to the normal distribution (also known as the Gaussian distribution or the bell curve) but has heavier tails. The shape of the t-distribution is determined by the degrees of freedom, which is closely related to the sample size (degrees of freedom = sample size - 1). As the degrees of freedom increase (i.e., as the sample size increases), the t-distribution approaches the normal distribution. [Formula of degrees of freedom]

In hypothesis testing and confidence interval estimation, the t-distribution is used in place of the normal distribution when the sample size is small (usually less than 30) and the population standard deviation is unknown. The t-distribution accounts for the additional uncertainty that arises from estimating the population standard deviation using the sample standard deviation.

To use the t-distribution in practice, you look up critical t-values from a t-distribution table, which provides values corresponding to specific degrees of freedom and confidence levels (e.g., 95% confidence). These critical t-values are then used to calculate confidence intervals or perform hypothesis tests.

Titanic Case Study

31 March 2023 18:00

Pop → 1360

$\mu \rightarrow \times$

$\sigma \rightarrow \times$

CLT → 10 times → 30 size

95% confidence level

inference