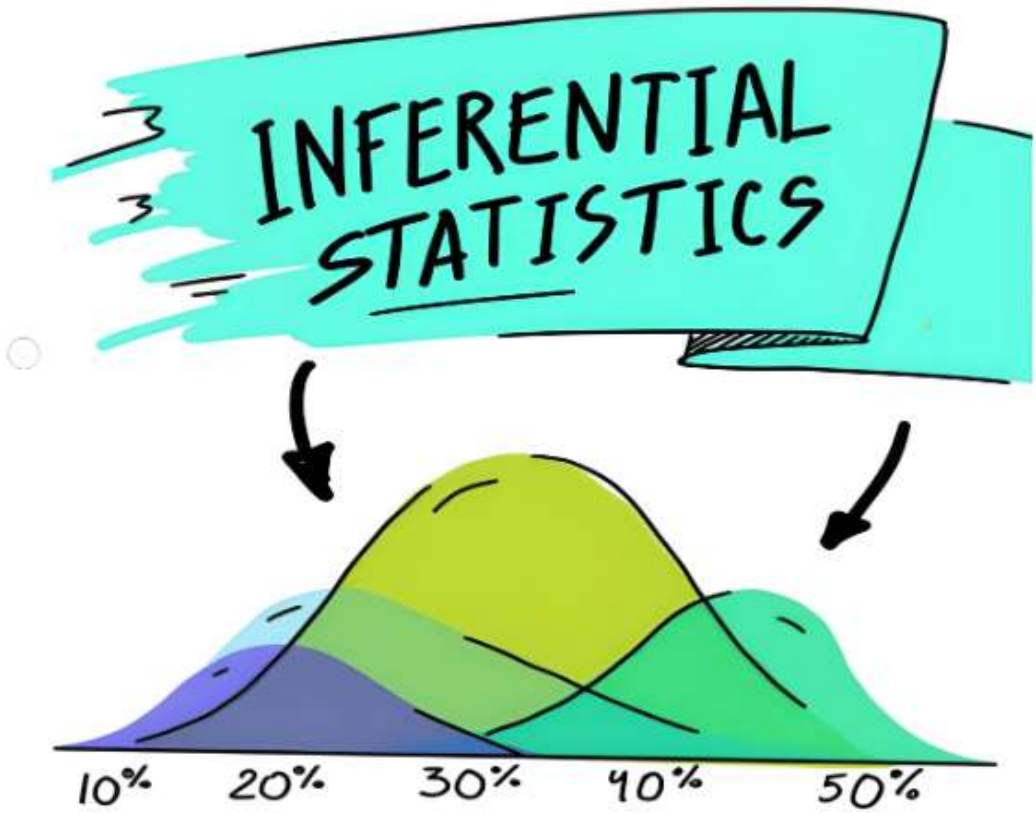
1. What is Inferential Statistics ?

- Inferential statistics deals with making conclusions and predictions about a **population based on a sample**. It involves the use of probability theory to estimate the likelihood of certain events occurring, hypothesis testing to determine if a certain claim about a population is supported by the data, and regression analysis to examine the relationships between variables



It involves using statistical techniques to test hypotheses and draw conclusions from data. Some of the topics that come under inferential statistics are:

1. **Hypothesis testing**: This involves testing a hypothesis about a population parameter based on a sample of data. For example, testing whether the mean height of a population is different from a given value.

2. **Confidence intervals**: This involves estimating the range of values that a population parameter could take based on a sample of data. For example, estimating the population mean height within a given confidence level.

3. **Analysis of variance (ANOVA)**: This involves comparing means across multiple groups to determine if there are any significant differences. For example, comparing the mean height of individuals from different regions.

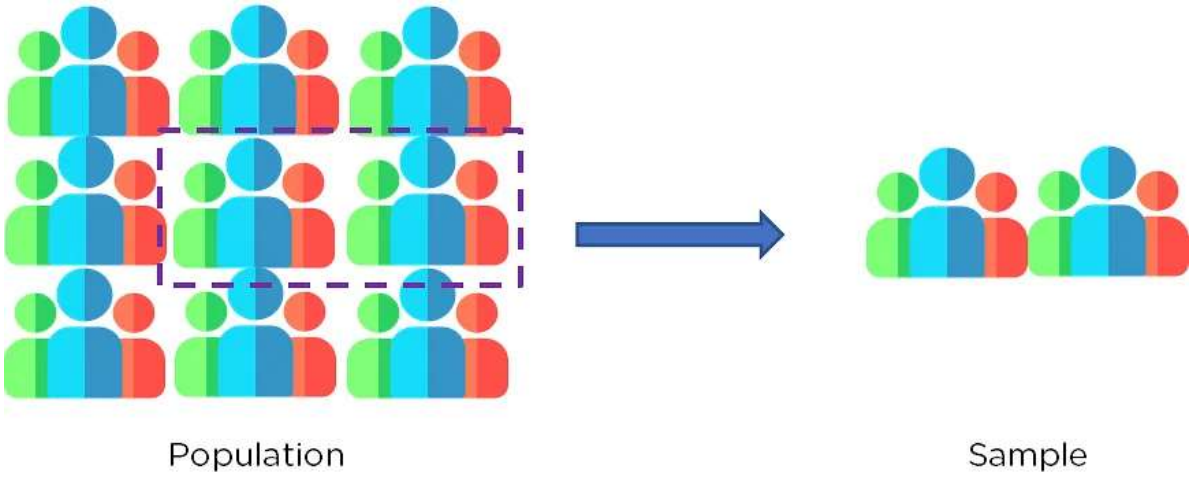
4. **Regression analysis**: This involves modelling the relationship between a dependent variable and one or more independent variables. For example, predicting the sales of a

product based on advertising expenditure.

5. **Chi-square tests**: This involves testing the independence or association between two categorical variables. For example, testing whether gender and occupation are independent variables.
6. **Sampling techniques**: This involves ensuring that the sample of data is representative of the population. For example, using random sampling to select individuals from a population.
7. **Bayesian statistics**: This is an alternative approach to statistical inference that involves updating beliefs about the probability of an event based on new evidence. For example, updating the probability of a disease given a positive test result

Some Terms

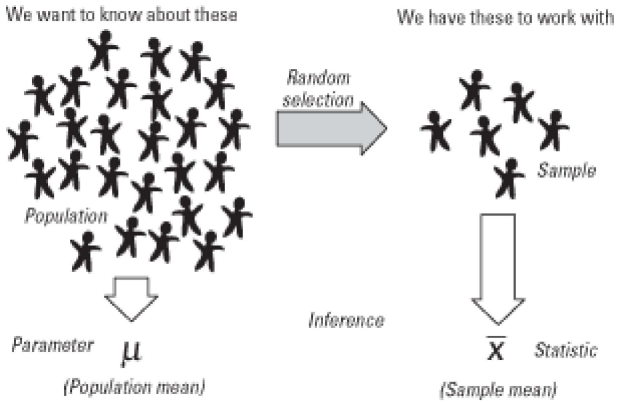
Population Vs Sample



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.

Parameter Vs Estimate(Statistic)



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.

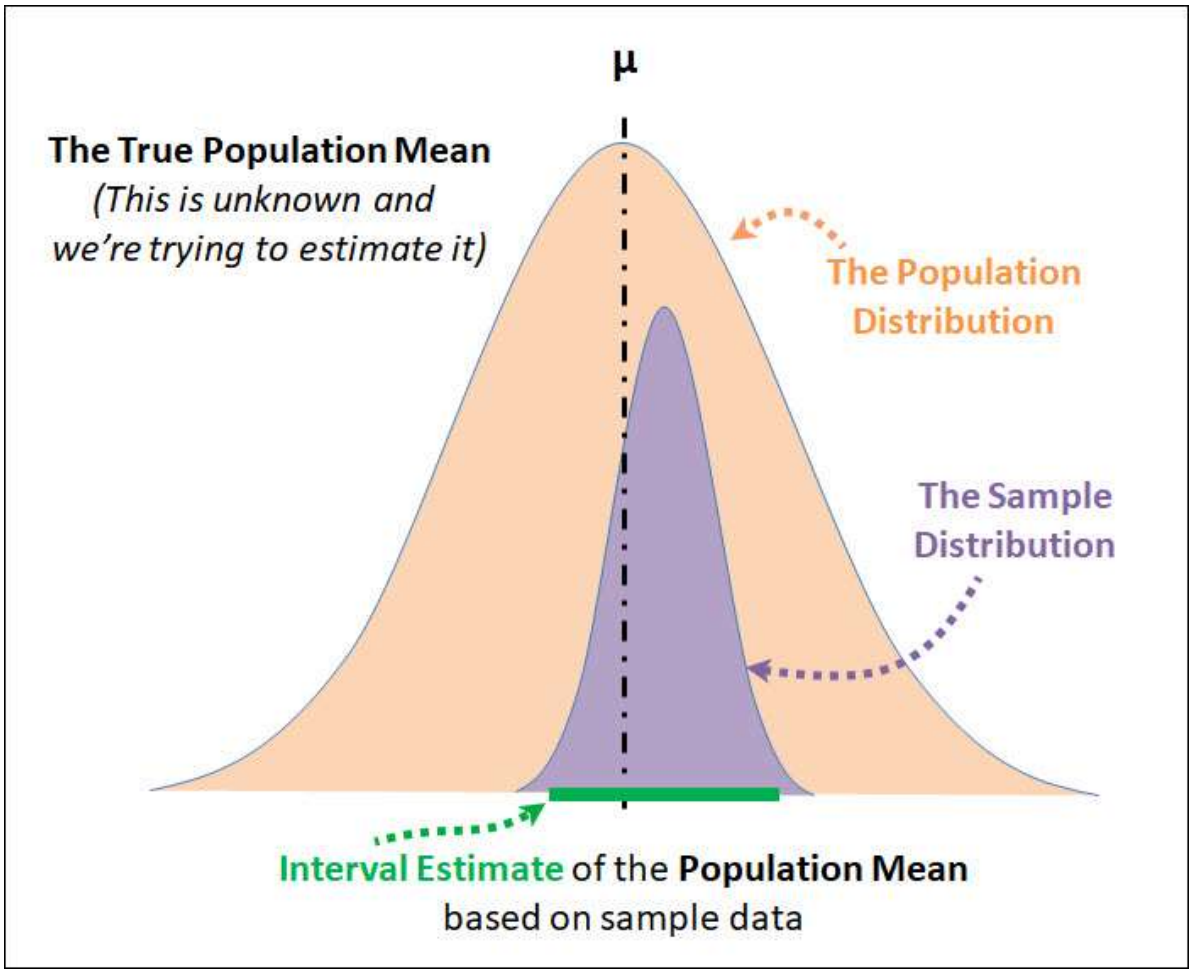
Statistic: A statistic is a numerical value that describes a characteristic of a sample, which is a subset of the population. By using statistics calculated from a representative sample, researchers can make inferences about the unknown respective parameter of the population. Common statistics include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

These methods help researchers answer questions like:

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

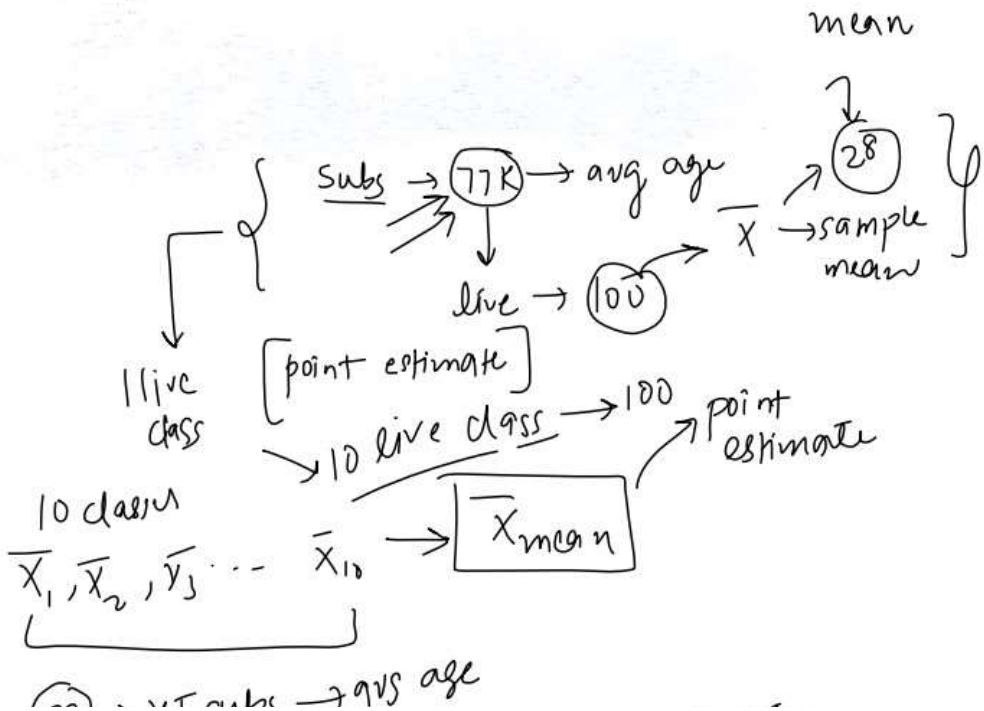
Inferential statistics are widely used in various fields, such as economics, social sciences, medicine, and natural sciences, to make informed decisions and guide policy based on limited data.

Point Estimate

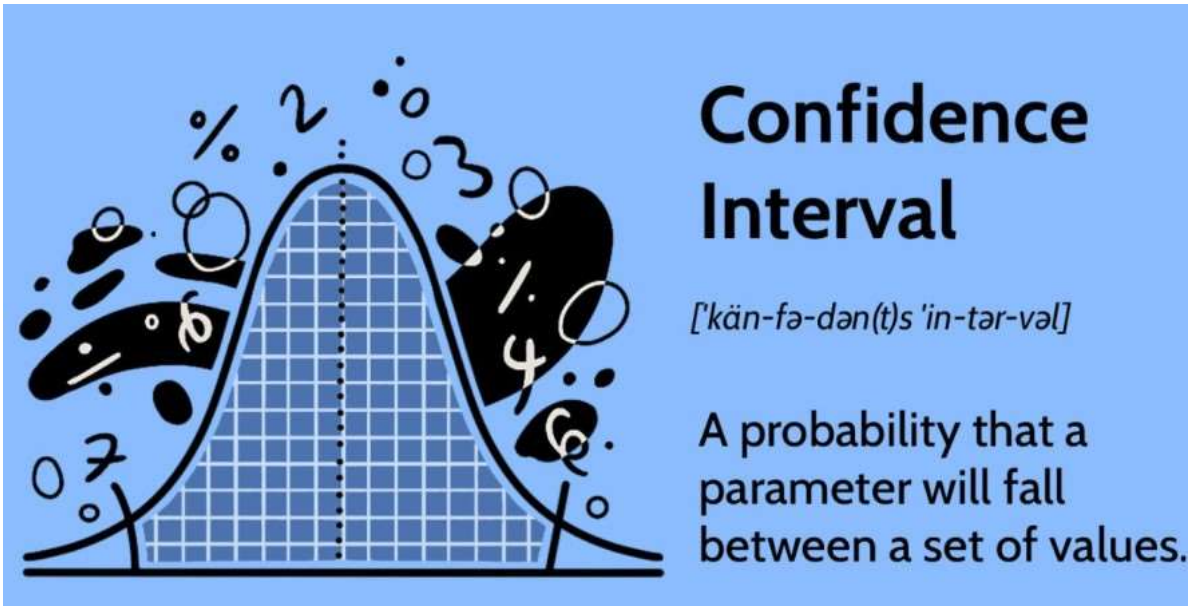


A **point estimate** is a single value, calculated from a sample, that serves as the best guess or approximation for an unknown population parameter, such as the mean or standard deviation. Point estimates are often used in statistics when we want to make inferences about a population based on a sample.

Explanation



Confidence Interval

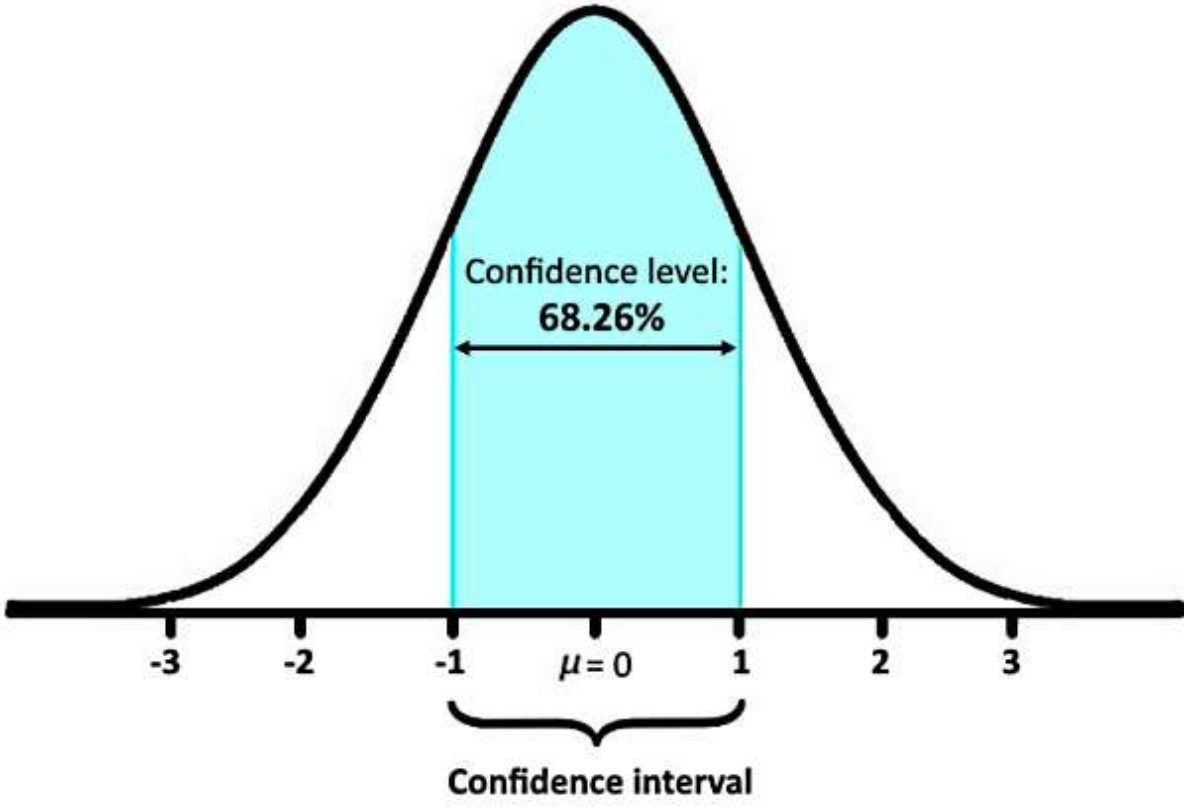


Confidence interval, in simple words, is a range of values within which we expect a particular population parameter, like a mean, to fall. It's a way to express the uncertainty around an estimate obtained from a sample of data.

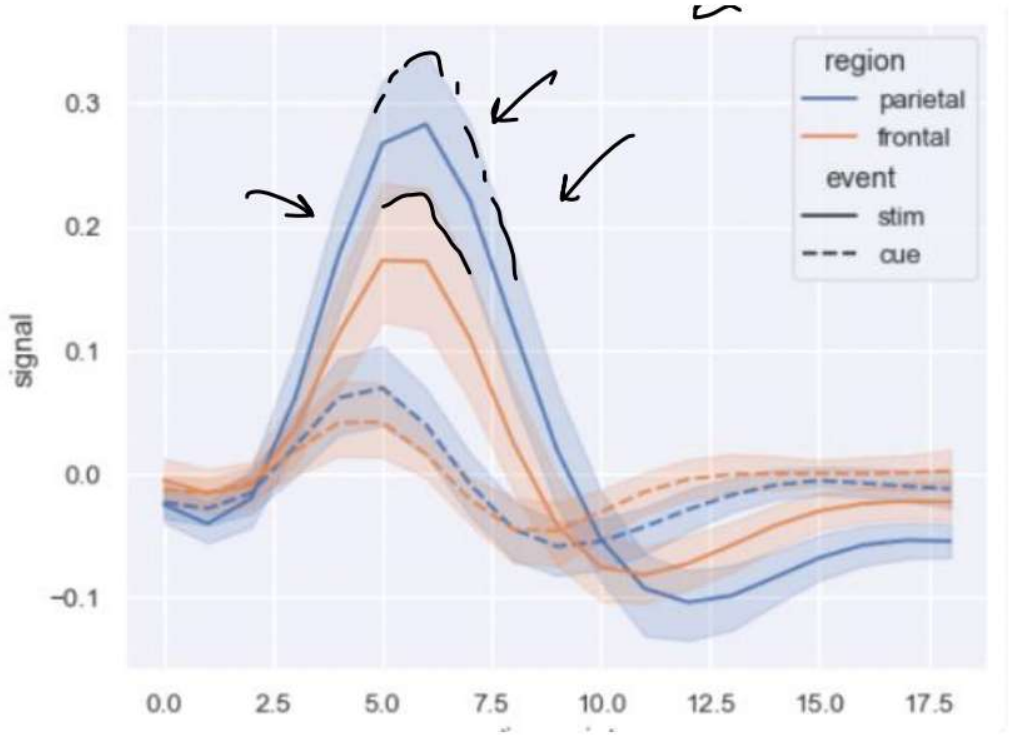
- **Confidence level**, usually expressed as a percentage like **95%**, indicates how sure we are that the true value lies within the interval.

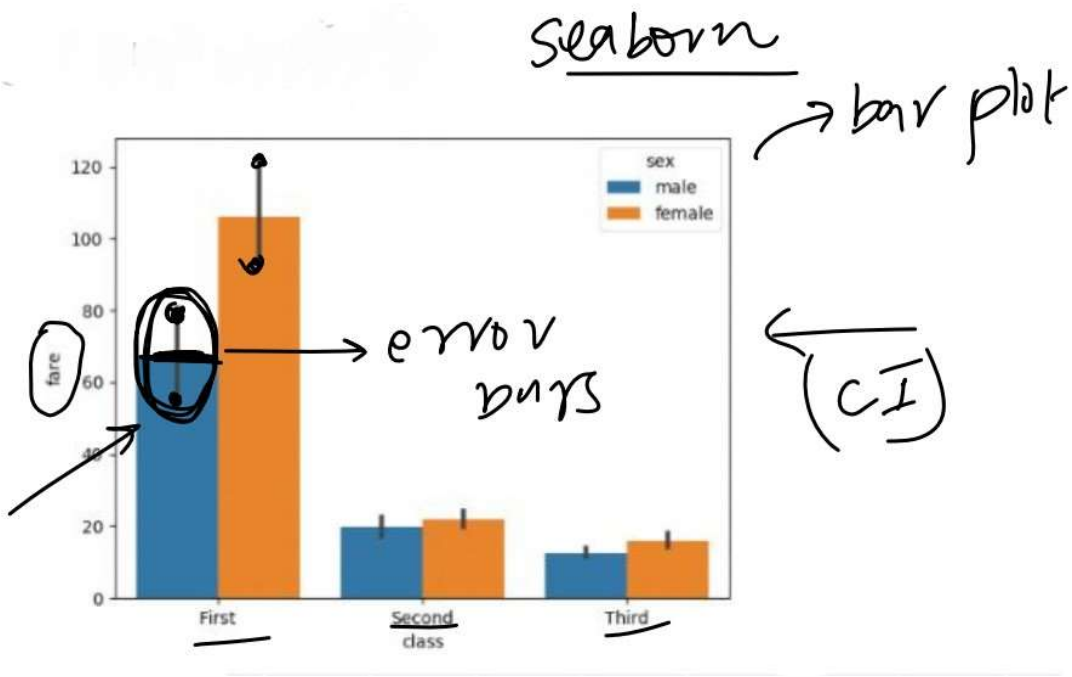
```
**Confidence Interval = Point Estimate + Margin of Error**
```

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



Examples of Confidence Interval





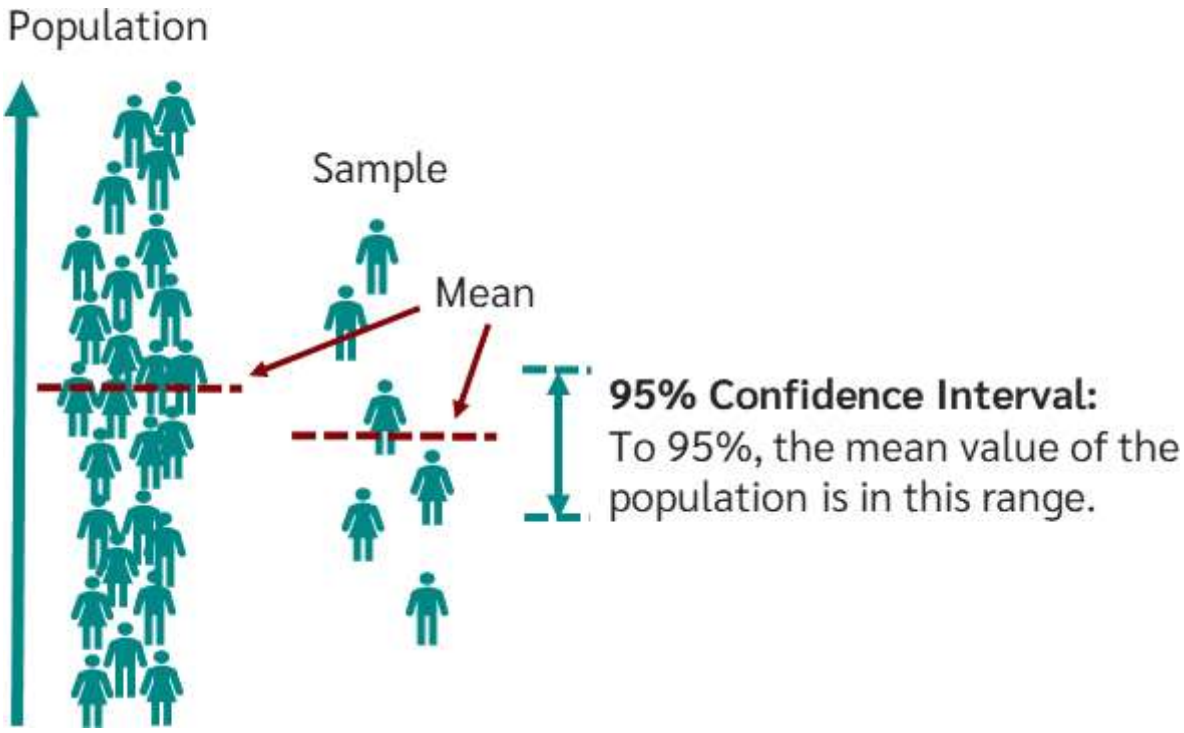
Confidence Interval (Sigma Known)

Assumptions

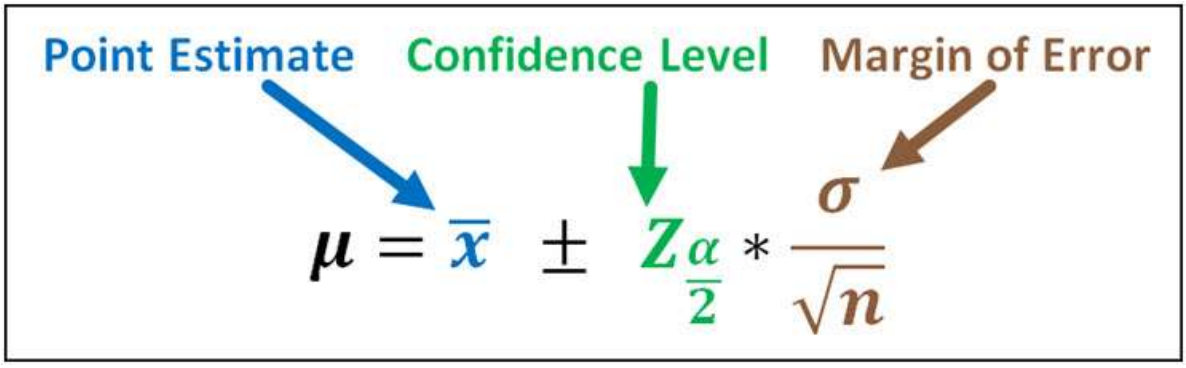
1. **Random sampling**: The data must be collected using a random sampling method to ensure that the sample is representative of the population. This helps to minimize biases and ensures that the results can be generalized to the entire population.

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.

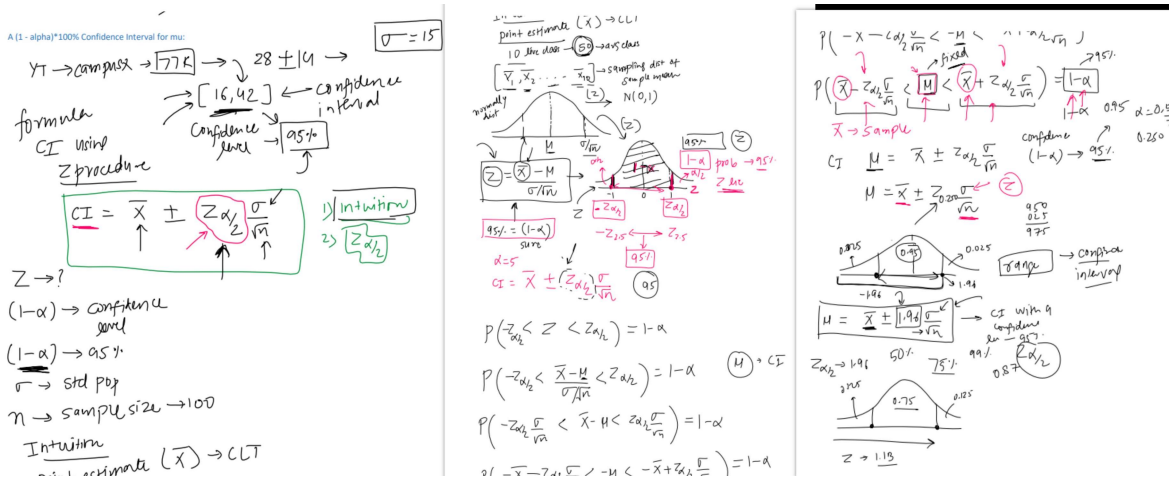
3. **Normal distribution or large sample size**: The Z-procedure assumes that the underlying population is normally distributed. However, if the population distribution is not normal, the Central Limit Theorem can be applied when the sample size is large (usually, sample size n ≥ 30 is considered large 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.



Formula



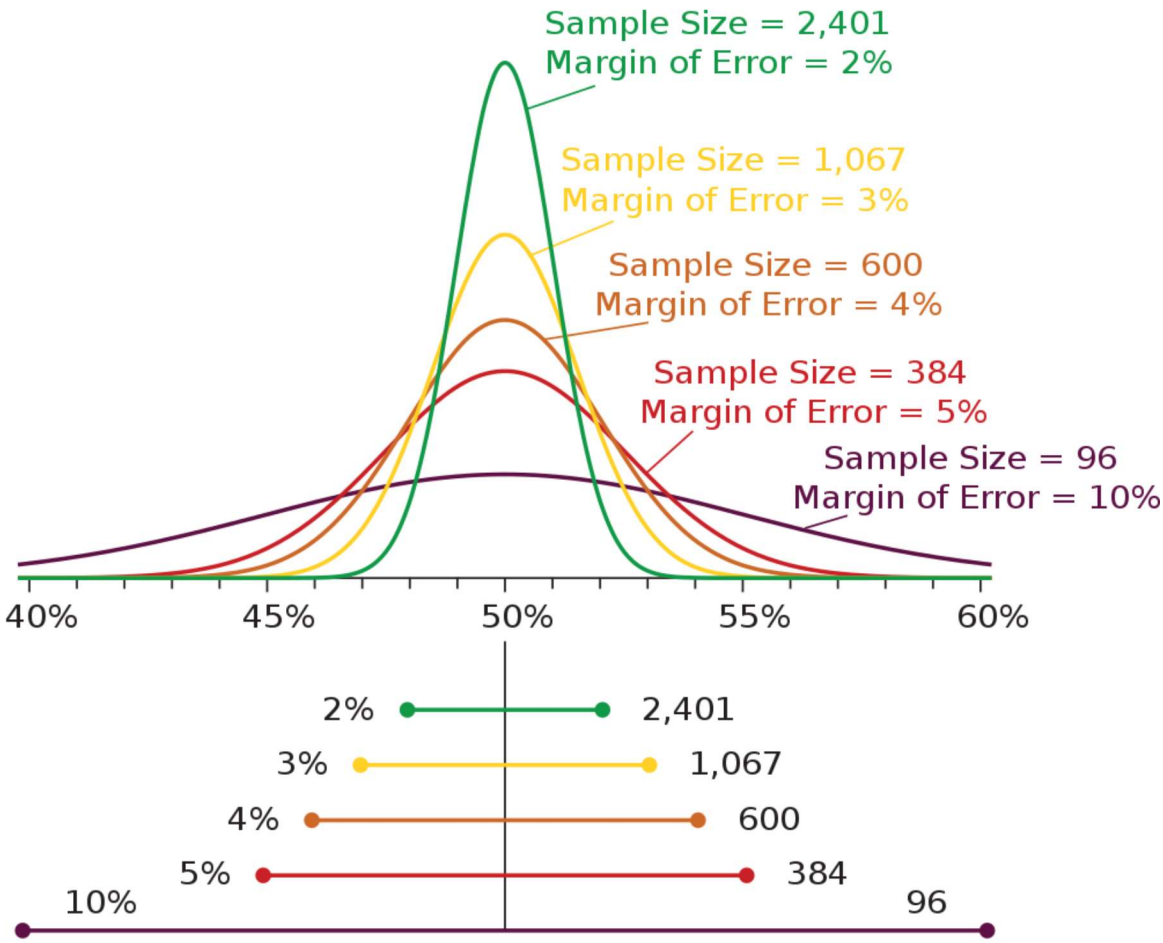
Explanation



What is a Margin of Error?

- **A margin of error tells you how many percentage points your results will differ from the real population value.**

  For example, a 95% confidence interval with a 4 percent margin of error means that your statistic will be within 4 percentage points of the real population value 95% of the time.



Formula

$$\mathrm{MoE} = z \times \sqrt{\frac{p(1-p)}{n}}$$

n = sample size

z = z-score

p = sample proportion

	A	B	C
1			
2	Mean	14	
3	Standard Deviation	2	
4	Sample Size	100	
5	Z-Score for 95%	1.96	
6			
7	**Margin of Error:**	=B5*(B3/SQRT(B4))	
8			
9			
10			
11			

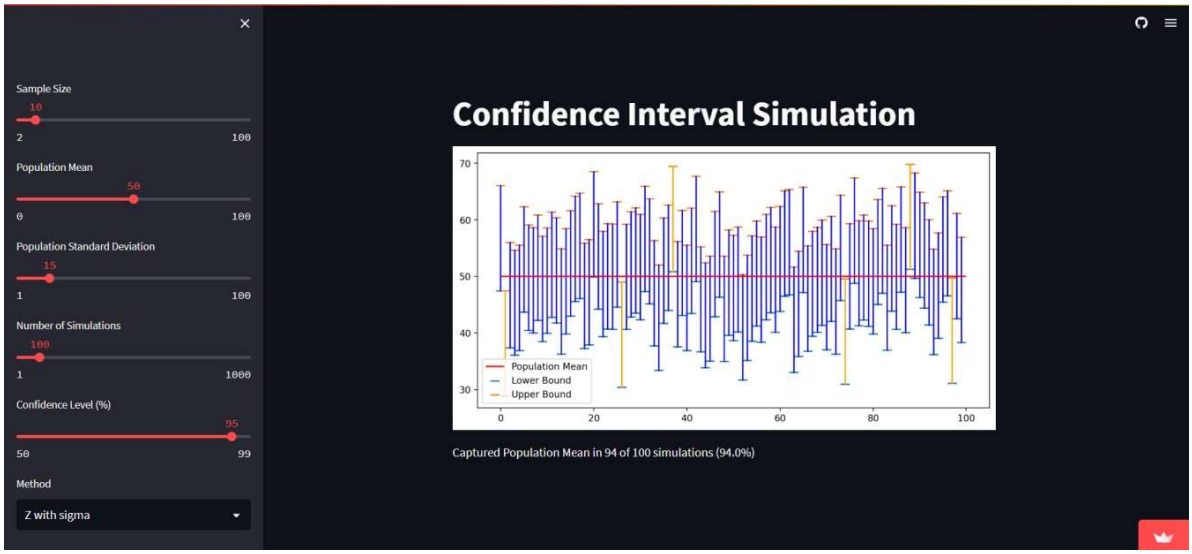
Interpreting Confidence Interval

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:

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.

https://campusx-official-confidence-interval-viz-app-kwg6wq.streamlit.app/ (https://campusx-official-confidence-interval-viz-app-kwg6wq.streamlit.app/)

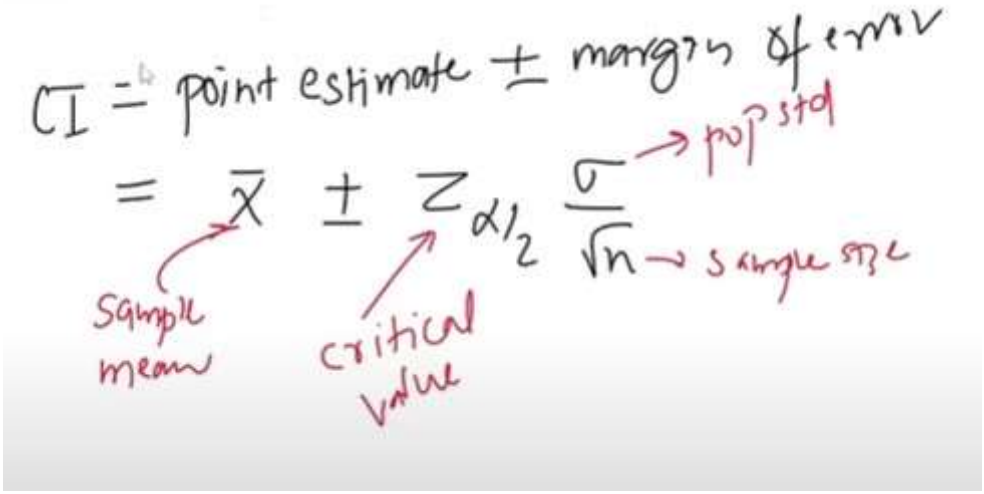


2. **Interval range**: The width of the confidence interval gives an indication of the precision of the estimate. A narrower confidence interval suggests a more precise estimate of the population parameter, while a wider interval indicates greater uncertainty. The width of the interval depends on the 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

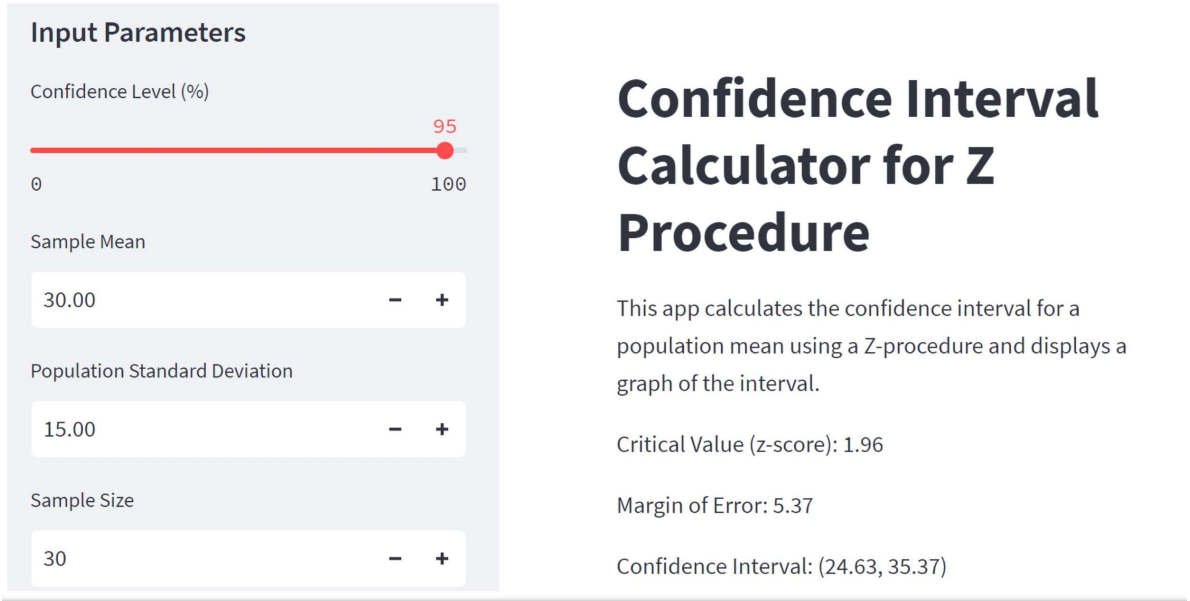
Factors Affecting Margin of Error

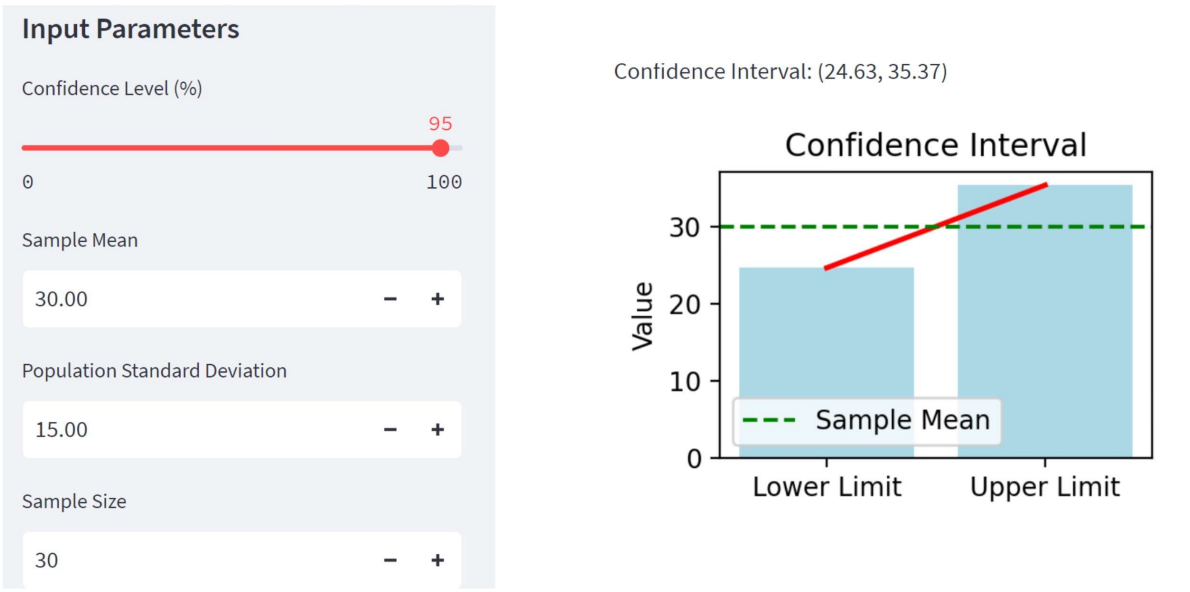
1. **Confidence Level (1-alpha)**
2. **Sample Size**
3. **Population Standard Deviation**



- Confidence Level (1-alpha)

https://campusx-official-z-distribution-conf-confidence-interval-bx6u60.streamlit.app/ (https://campusx-official-z-distribution-conf-confidence-interval-bx6u60.streamlit.app/)





Margin of Error and Population Standard Deviation (Z Procedure)

In [3]:
```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

# Set parameters
sample_size = 30
confidence_level = 95

# Calculate the critical value for the Z-score
critical_value = stats.norm.ppf((1 + confidence_level / 100) / 2)

# Generate an array of population standard deviations
population_std_devs = np.arange(1, 101)

# Calculate the margin of error for each population standard deviation
margin_of_errors = critical_value * (population_std_devs / np.sqrt(sample_size

# Plot the relationship
plt.plot(population_std_devs, margin_of_errors)
plt.xlabel("Population Standard Deviation")
plt.ylabel("Margin of Error")
plt.title("Relationship Between Margin of Error and Population Standard Deviat
plt.grid(True)
plt.show()
```

Margin of Error and Sample Size

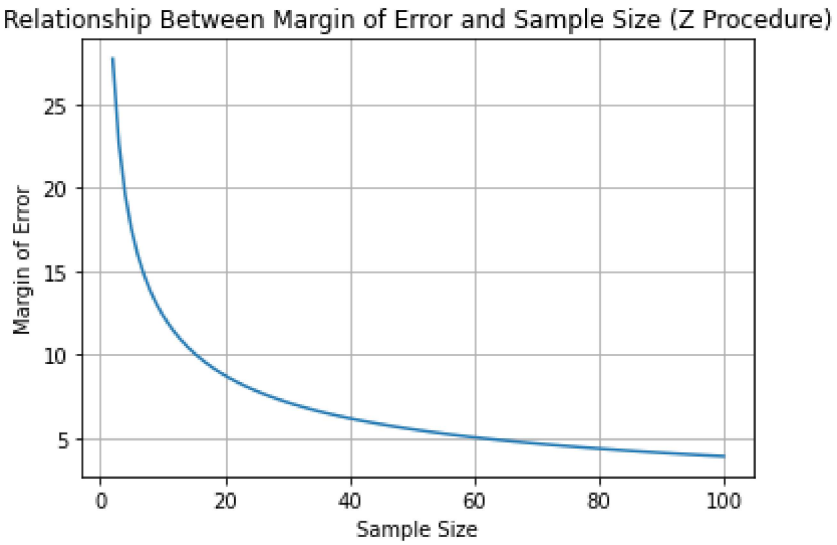
In [4]:
```
# Set parameters
population_std_dev = 20
confidence_level = 95

# Calculate the critical value for the Z-score
critical_value = stats.norm.ppf((1 + confidence_level / 100) / 2)

# Generate an array of sample sizes
sample_sizes = np.arange(2, 101)

# Calculate the margin of error for each sample size
margin_of_errors = critical_value * (population_std_dev / np.sqrt(sample_sizes

# Plot the relationship
plt.plot(sample_sizes, margin_of_errors)
plt.xlabel("Sample Size")
plt.ylabel("Margin of Error")
plt.title("Relationship Between Margin of Error and Sample Size (Z Procedure)"
plt.grid(True)
plt.show()
```



Margin of Error and Critical Value (Z Procedure)

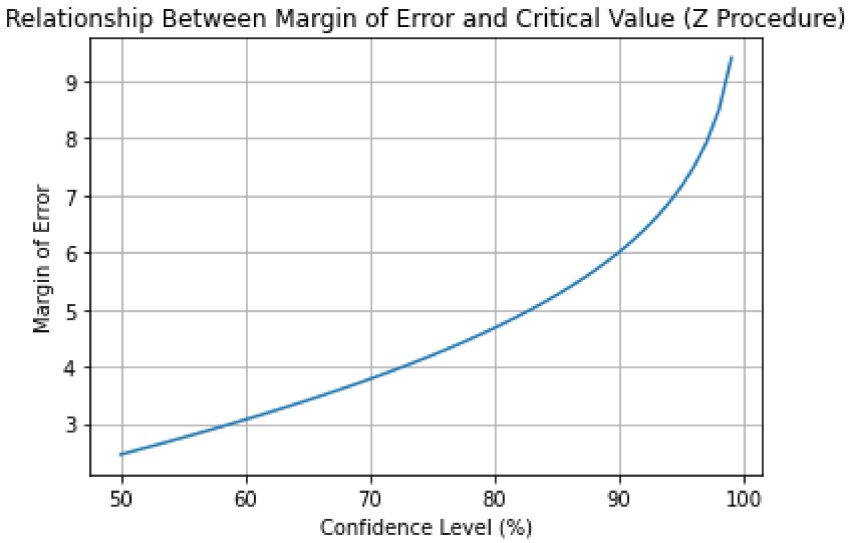
In [5]:
```
# Set parameters
sample_size = 30
population_std_dev = 20

# Generate an array of confidence levels from 50% to 99%
confidence_levels = np.arange(50, 100)

# Calculate the critical values (Z-scores) for each confidence level
critical_values = stats.norm.ppf((1 + confidence_levels / 100) / 2)

# Calculate the margin of error for each critical value
margin_of_errors = critical_values * (population_std_dev / np.sqrt(sample_size

# Plot the relationship
plt.plot(confidence_levels, margin_of_errors)
plt.xlabel("Confidence Level (%)")
plt.ylabel("Margin of Error")
plt.title("Relationship Between Margin of Error and Critical Value (Z Procedur
plt.grid(True)
plt.show()
```



Confidence Interval (Sigma not known)

Using the t procedure

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.

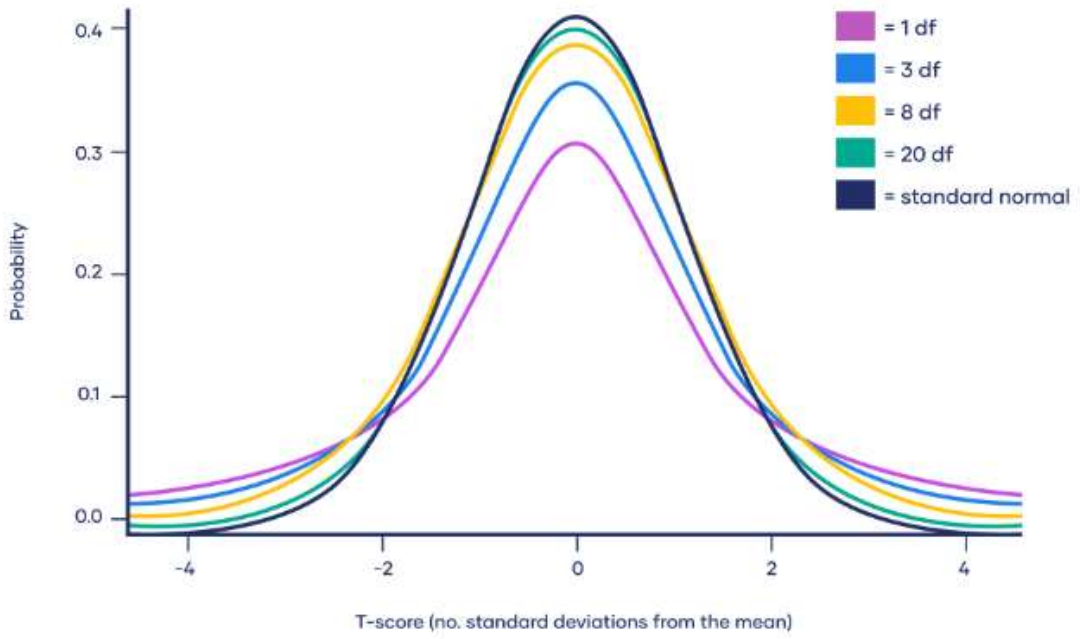
2.**Sample standard deviation**: The population standard deviation (σ) is unknown, and the sample standard deviation (s) is used as an estimate. The t-distribution is specifically designed to account for the additional uncertainty introduced by using the sample standard deviation

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

Student's T Distribution



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."

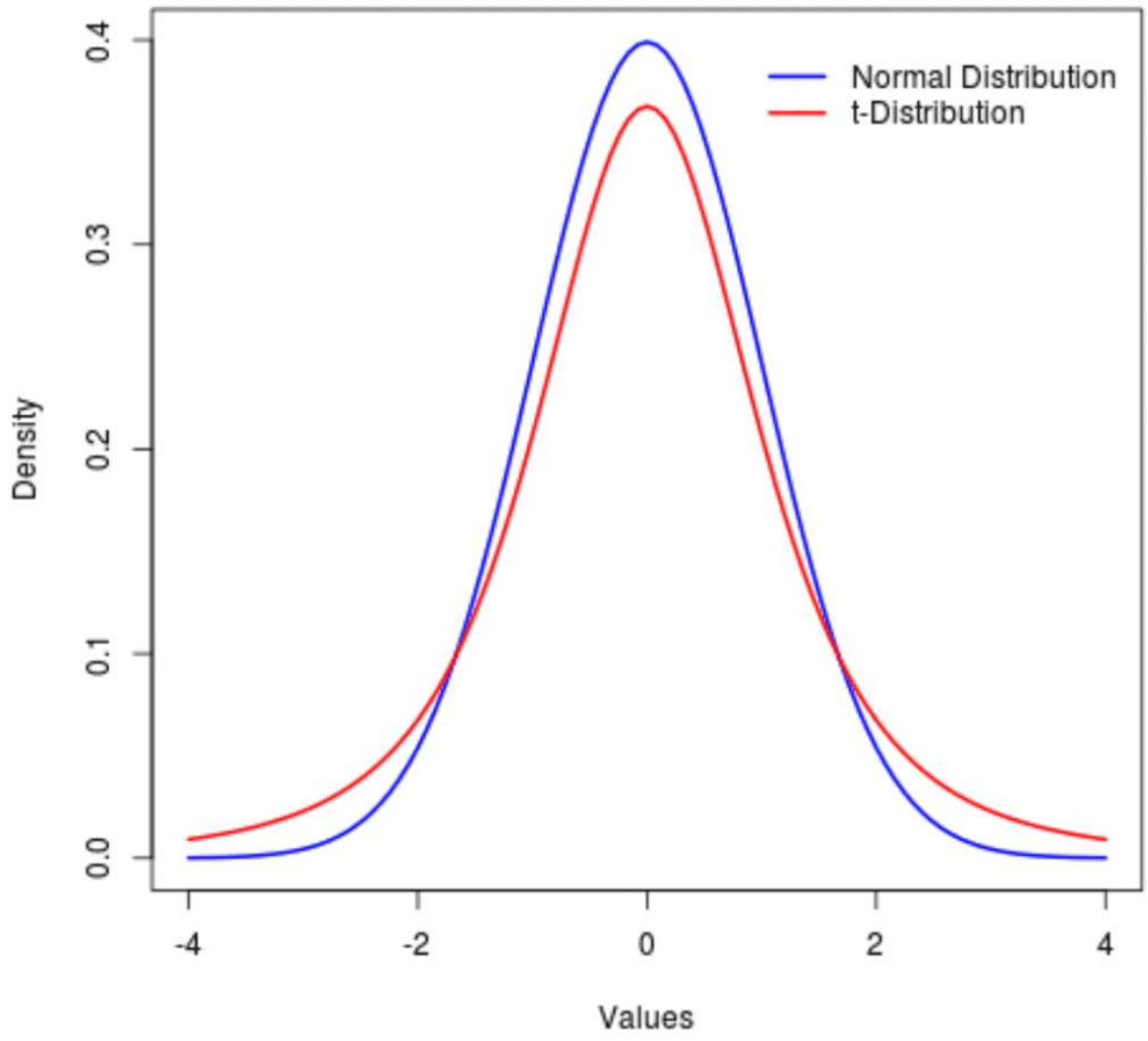
- The t-distribution is similar to the normal distribution (also known as the Gaussian distribution or the bell curve) but has heavier tails. The shape of the t-distribution is determined by the degrees of freedom, which is closely related to the sample size (degrees of freedom = sample size - 1). As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

- 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

https://campusx-official-normal-distribution-vs-t-distributi-app-28si1q.streamlit.app/ (https://campusx-official-normal-distribution-vs-t-distributi-app-28si1q.streamlit.app/)



T-table

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

What's the key difference between the t- and z-distributions?

- The standard normal or **z-distribution** assumes that you know the **population standard deviation**. The **t-distribution** is based on the **sample standard deviation**.

t-distribution Formula

Formula

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

t = Student's t-distribution

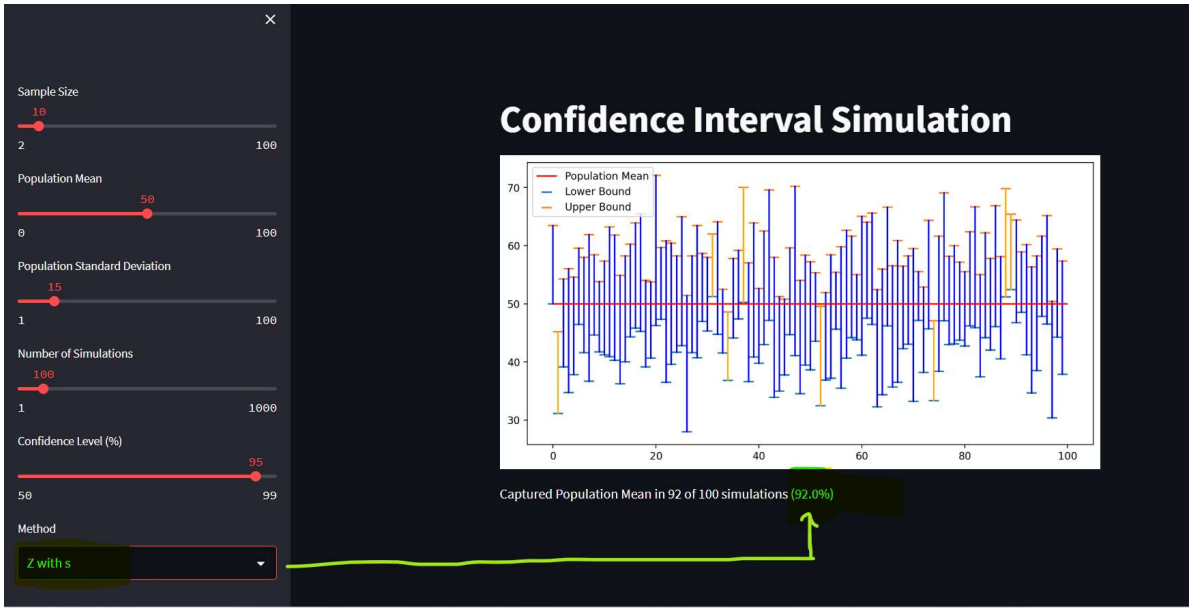
$\bar{x}$ = sample mean

μ = population mean

s = sample standard deviation

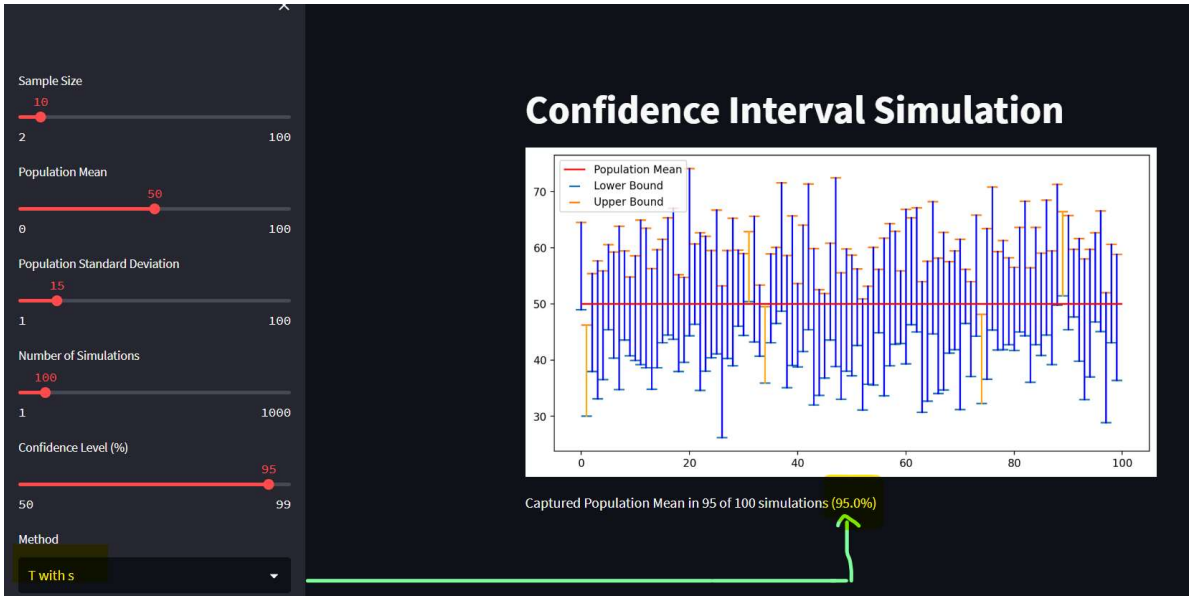
n = sample size

- What if we use Z Statistic with sample



- it gives Confidence level as 92% not 95% - Because its Not correct statistic , Always 'Z' is calculated with Population, not with sample

- What if we use T distibution with sample



- it gives Confidence level as Exaxt 95% - Because its correct statistic

T distribution code

In [11]:
```
import pandas as pd
import numpy as np
```

In [12]:
```
train_df = pd.read_csv("C:\\Users\\user\\Downloads\\train.csv")
test_df = pd.read_csv("C:\\Users\\user\\Downloads\\test.csv")
```

In [14]:
```
train_df.shape
```

Out[14]: (891, 12)

In [15]:
```
test_df.shape
```

Out[15]: (418, 11)

In [16]:
```
# Concat
df = pd.concat([train_df.drop(columns=['Survived']),test_df]).sample(1309)
```

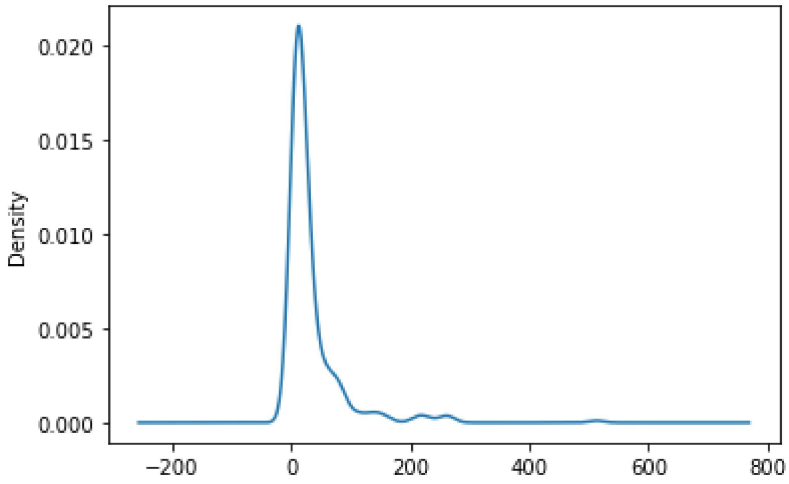
In [18]:
```
df.shape # 1309
```

Out[18]: (1309, 11)

In [19]:
```
# On Fare column

df['Fare'].plot(kind='kde')
```

Out[19]: <AxesSubplot:ylabel='Density'>



In [53]:
```
# sample size = 30 -> 10 times

""" For each iteration ,
we are taking out the samples and also calculating Standard deviation """

samples = []
stds=[]

for i in range(10):
    x = df['Fare'].dropna().sample(30).values
    stds.append(x.std())
    samples.append(x.tolist())
```

In [54]:
```
# Converting Values to Numpy array

samples =np.array(samples)
```

In [55]:
```
# X-bar

sampling_means =samples.mean(axis=1)
```

In [56]:
```
# std
sample_std =np.mean(stds)
```

In [41]:
```
sample_std
```

Out[41]: 54.57891877236538

In [57]:
```
# n
sampling_means.std()/np.sqrt(30)
```

Out[57]: 1.0556248313286112

With 95% Confidence interval

27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326
	0%	50%	60%	70%	80%	90%	95%	98%
					Confidence Level			

- For 30 samples With 95% Confidence Level , we got 2.042

In [58]:
```
# with 95% Confidence

lower_limit = sampling_means.mean() - 2.042*(sample_std/np.sqrt(30))
upper_limit = sampling_means.mean() + 2.042*(sample_std/np.sqrt(30))
```

In [59]:
```
lower_limit
```

Out[59]: 14.386461391934132

In [60]:
```
upper_limit
```

Out[60]: 43.44162060806586

In [61]:
```
print('The range is',lower_limit,'-',upper_limit)
```

```
The range is 14.386461391934132 - 43.44162060806586
```

In [62]:
```
df['Fare'].mean()
```

Out[62]: 33.29547928134563

With 50% Confidence interval

28	0.000	0.683	0.855	1.056	1.313
29	0.000	0.683	0.854	1.055	1.311
30	0.000	0.683	0.854	1.055	1.310
40	0.000	0.681	0.851	1.050	1.303
60	0.000	0.679	0.848	1.045	1.296
80	0.000	0.678	0.846	1.043	1.292
100	0.000	0.677	0.845	1.042	1.290
1000	0.000	0.675	0.842	1.037	1.282
z	0.000	0.674	0.842	1.036	1.282
	0%	50%	60%	70%	80%

- For 30 samples With 50% Confidence Level , we got 0.683

In [63]:
```
# with 50% Confidence

lower_limit = sampling_means.mean() - 0.683*(sample_std/np.sqrt(30))
upper_limit = sampling_means.mean() + 0.683*(sample_std/np.sqrt(30))
```

In [64]:
```
lower_limit
```

Out[64]:
```
24.054914226097456
```

In [65]:
```
upper_limit
```

Out[65]:
```
33.773167773902536
```

In [66]:
```
print('The range is',lower_limit,'-',upper_limit)
```

```
The range is 24.054914226097456 - 33.773167773902536
```

In [ ]:
```

```